Recuperación de pasajes en textos legales y patentes multilingües

Santiago Correa García

Departamento de Sistemas Informáticos y Computación

Director: Paolo Rosso

Tesis desarrollada dentro del Máster en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

Valencia, Julio de 2010

Resumen

La búsqueda de información específica, que satisfaga las necesidades de un usuario en una extensa colección de documentos, es una tarea dispendiosa que demanda una gran cantidad de tiempo, y en la mayoría de los casos expertos en la materia que interpreten la información. Los altos costos que dicha tarea representa, para sectores como el gubernamental o el empresarial, han generado altas expectativas en el desarrollo de sistemas automáticos, que reduzcan el tiempo de respuesta y los costos asociados al proceso. La *Recuperación de Pasajes en Textos Legales y Patentes Multilingües*, tal y como se plantea en la presente investigación, es un proceso computacional que intenta solucionar la problemática planteada anteriormente.

En este trabajo se exponen tres temas principales: la problemática de la *Recuperación de Pasajes*, el dominio de los *Textos Legales y las Patentes*, y la característica de la *diversidad idiomática*; los cuales en pocas ocasiones son abordados de manera conjunta. En el documento se presenta la descripción del estado del arte en materia de *Recuperación de Pasajes* y el procesamiento de *Textos Legales y Patentes*. Además, se presentan técnicas que permitan abordar problemas de *Recuperación de Información*, que involucren los tres temas principales. Por último se analizan dos participaciones en competencias que pretenden evaluar el desempeño de los enfoques planteados.

Como continuación de los experimentos realizados, se plantean posibles líneas de investigación que podrían ser tratadas en un posible estudio doctoral.

Agradecimientos

Quisiera agradecer en primer lugar al profesor Paolo Rosso, por la supervisión del presente trabajo, así como a Davide Buscaldi, por la ayuda brindada en la competencia de búsqueda de respuesta y propiedad intelectual del CLEF.

En segundo lugar quisiera agradecer a la compañía Maat Gknowledge, por la beca de colaboración con la Universidad Politécnica de Valencia, en el contexto del proyecto: "Modulo de servicios semánticos de la plataforma G".

Esta tesis ha sido desarrollada en el marco del proyecto MICINN (Plan I+D+i): TEXT-ENTERPRISE 2.0, Técnicas de Comprensión de textos aplicadas a las necesidades de la Empresa 2.0. (TIN2009-13391-C04-03).

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

La vertiginosa aceleración del consumo de los mercados a nivel mundial, genera voraces espacios de competencia, en los que la demanda de información oportuna en un corto periodo de tiempo, se ha convertido en una de las herramientas de mayor eficacia para la supervivencia de las empresas [58]. Por tal motivo, numerosas investigaciones a nivel mundial, proponen técnicas que intentan satisfacer las necesidades de búsqueda de información, relevante a las expectativas de los distintos usuarios en temas tan diversos como: economía, leyes, ciencias, computación, tendencias de mercados, etc. o en simples búsquedas a nivel del público en general, expresadas como una pregunta puntual: *"¿Cuál es la capital de Madagascar?"* o *"¿A cuántos centímetros equivale una pulgada?"*.

La presente tesis, aborda el tema de la recuperación de pasajes en textos legales y patentes multilingües, enfatizando la importancia del tema en la cotidianidad y el papel de la investigación en el logro de resultados, que permitan la consecución de técnicas eficaces para su utilización en escenarios reales.

1.1. Descripción de la problemática

En los últimos años se ha experimentado un creciente interés científico en el área de procesamientos de textos. Las grandes colecciones de documentos disponibles, demandan de los investigadores el desarrollo de técnicas computacionales, que permitan la *Recuperación de Información* (*RI*) de utilidad en la empresa. Para promover el adecuado desarrollo y la comparación de las distintas técnicas, varios foros a nivel mundial generan los espacios necesarios para la investigación de dichas actividades. En la Tabla 1.1 pueden ser apreciados tres importantes foros que promueven la investigación en técnologias de acceso a la información.

Tabla 1.1: Foros que promueven la investigación en tecnologías de acceso a la información

Nombre	Sigla	Región
NII-NACSIS Test Collection for IR Systems	*NTCIR*	Asia
The Text Retrieval Conference	*TREC*	América
The Cross-Language Evaluation Forum	*CLEF*	Europa

En las tareas propuestas por los foros: *NTCIR*[1], *TREC*[2] y *CLEF*[3] para incentivar la investigación se desarrollan entre otros, los siguientes temas: *Recuperación de Información*, *Búsqueda de Respuesta*, *Resumen de Textos* y *Extracción de Información*. Algunas de estas tareas son impulsadas por entidades reconocidas en el área de *Recuperación de Información*, como es el caso de:

- *MatrixWare*[4]: Firma que ofrece soluciones y servicios de recuperación de información profesional para el mercado global, proporcionando ventajas competitivas a sus clientes.
- *IRF*[5]: Institución sin ánimo de lucro que tiene como objetivo promover y facilitar la investigación en la *Recuperación de Información*.

En el desarrollo de los Capítulos 4 y 5 se estudia con mayor detalle algunas de las competencias promovidas por el foro Europeo *CLEF*.

Es importante destacar que, el presente documento centra su interés en las competencias de *Recuperación de Información* y *Búsqueda de Respuesta* aplicadas al dominio legal, con el ánimo de dar continuidad y explorar los alcances de los enfoques desarrollados por el laboratorio *NLEL*[6] (Natural Language Engineering Lab) del grupo de investigación *ELiRF*[7] (Grupo de Ingeniería del Lenguaje Natural y Reconocimiento de Formas) de la *Universidad Politécnica de Valencia* a lo largo de los últimos años, en el área de *Procesamiento del Lenguaje Natural* (*PLN*).

Como ya se ha mencionado, las grandes colecciones de texto de las que dispone la universidad generan la necesidad de ser procesadas eficientemente, para extraer de sí información relevante a las necesidades de los usuarios. Entre las colecciones de texto de mayor interés por su potencial utilidad, están las colecciones de carácter legal, ejemplos de ellas son:

- Europarl[8]: La colección de documentos se extrae de las actas del *Parlamento Europeo*. Cada documento incluido en la colección está traducido en 11 idiomas europeos:

[1] http://research.nii.ac.jp/ntcir/
[2] http://trec.nist.gov/
[3] http://www.clef-campaign.org/
[4] http://www.matrixware.com/
[5] http://www.ir-facility.org/
[6] http://users.dsic.upv.es/grupos/nle/
[7] http://elirf.dsic.upv.es/elirf/
[8] http://www.statmt.org/europarl/

Francés, Italiano, Español, Portugués, Inglés, Neerlandés, Alemán, Danés, Sueco, Griego y Finlandés.

- Hansard[9]: Colección de documentos extraídos de los debates del *Parlamento Canadiense*, publicados en los dos idiomas oficiales del país: Inglés y Francés.

- JRC-Acquis[10]: Totalidad de las leyes aplicadas en los estados miembros de la Unión Europea, la colección de datos JRC-Acquis se estudia más detalladamente en la Sección 5.3.1.

La activa participación de los distintos grupos de investigación a nivel mundial en el desarrollo de las distintas competencias, ha posibilitado la generación de técnicas capaces de extraer información de extensas colecciones de documentos de manera eficiente, acercando el ámbito académico a las necesidades empresariales.

1.2. Motivación

La gran cantidad de información presente en la empresa, la universidad, o simplemente a nivel del público en general presente en la red, se encuentra generalmente escrita en el lenguaje con el cual los seres humanos se comunican o, en otras palabras *Lenguaje Natural* (*LN*). Para "dar un paso más" en la *Recuperación de la Información*, a través de ordenadores, la cual está tradicionalmente basada en la presencia de los términos de una consulta en los documentos analizados, se pretende dotar a los motores de búsqueda con un nivel de inteligencia superior, en el cual se analice la información suministrada por el usuario en *Lenguaje Natural*, para posteriormente procesar los datos de manera tal que sea retornada la mínima información necesaria que satisfaga sus necesidades, ya sea dando una respuesta exacta o suministrando fragmentos de texto (pasajes) extraídos de un documento, en el cual se encuentre la respuesta a la pregunta planteada por el usuario.

Los altos costos asumidos por las empresas, debidos al gran volúmen de datos que se procesan en tareas que exigen conocimientos específicos en un área determinada, como por ejemplo: el departamento legal de una compañía o el estudio de tendencias de mercados basado en los comentarios de los compradores [81, 48], han dado gran importancia al procesamiento automático de datos. Hoy en día, los recursos económicos derivados de la empresa, urgida de una reducción de costos en el ejercicio de sus actividades, potencia las investigaciones de dicha área.

En un intento por facilitar la transferencia de conocimiento entre la universidad y el mundo empresarial, en el desarrollo de la presente investigación se pretende cumplir con los

[9] http://www.isi.edu/natural-language/download/hansard/
[10] http://langtech.jrc.it/JRC-Acquis.html/

siguientes objetivos:

- Objetivos generales:
  1. Estudiar el alcance potencial de las herramientas de *Procesamiento de Lenguaje Natural* para su aplicación en problemas de la vida real, más específicamente en el dominio legal;
  2. Continuar el desarrollo de los avances del laboratorio *NLEL* en el desempeño de competencias de *Recuperación de Información*;
  3. Establecer las bases para la implementación de enfoques válidos para el desarrollo de tareas de carácter multilingüe.
- Objetivos específicos:
  1. Estudiar del estado del arte del procesamiento de textos legales y patentes;
  2. Revisar el estado del arte de la recuperación de pasajes;
  3. Generar experimentos que apliquen técnicas de *Recuperación de Información* para evaluar su desempeño en tareas reales;
  4. Diseñar enfoques basados en herramientas que han demostrado tener buenos resultados en tareas de *Recuperación de Información*, para su aplicación en el dominio legal de carácter multilingüe.

1.3. Planteamiento del problema

En el marco de la investigación realizada para el desarrollo de la presente tesis, varios experimentos en los que se han planteado enfoques basados en la recuperación de pasajes se han implementado para la participación del laboratorio *NLEL* en dos tareas de *Recuperación de Información* en el ámbito legal: *ResPubliQA 2009* e *IP 2009*. A continuación se hace una breve descripción de las tareas, las cuales posteriormente serán tratadas a profundidad en los Capítulos 4 y 5 respectivamente.

- ***ResPubliQA 2009***: Competencia de *Búsqueda de Respuesta*, en la cual se plantea el hipotético escenario en el que un usuario, por medio de preguntas formuladas en *Lenguaje Natural*, muestra su interés en conocer información concreta, relativa al campo legislativo, más específicamente a la legislación europea. Para dicha labor los sistemas participantes deben buscar las respuestas a cada pregunta formulada en un conjunto de documentos de legislación europea (*JRC-Acquis*[11]) que cuenta con traducciones paralelas, alineadas a nivel de documento en varias lenguas europeas [59].

[11]La colección de documentos *JRC-Acquis* será estudiada extensamente en la Sección 4.3.1

- ***IP 2009***: El objetivo de la competencia es la investigación de técnicas de *Recuperación de Información* en el dominio de la *Propiedad Intelectual* de las patentes. Uno de los requisitos principales para que una patente sea concedida, es el hecho de que dicha "posible" patente cumpla con el requisito de ser novedosa; es decir, no deben existir patentes anteriores u otra publicación que describa la invención. La publicación de un documento que entre en conflicto con una potencial idea de patente puede darse en cualquier lugar y en cualquier idioma; por tal motivo, cuando una persona realiza una búsqueda, ya sea para determinar si una idea es potencialmente patentable, o para tratar de probar que una patente no debería haberse concedido, dicha persona se enfrenta a una búsqueda inherentemente translingüe.

  La meta de la competencia consiste en que, dada una patente, encontrar un conjunto de patentes que estén relacionadas entre sí, o lo que es lo mismo, encontrar su estado del arte. Una gran colección de patentes (*EPO*[12]) escritas en tres idiomas - Inglés, Francés y Alemán - es suministrada a los participantes para la realización de la tarea [70].

1.4. Organización de la tesis

Los capítulos que componen la tesis, además del presente capítulo, son descritos brevemente a continuación:

Capítulo 2, **Procesamiento de textos legales y patentes**: En este capítulo se ofrece una visión general sobre el procesamiento de textos en el ámbito legal y de patentes, basándose en un marco de referencia dado por el estado del arte de cada tema.

Capítulo 3, **Recuperación de pasajes en textos multilingües**: Capítulo en el que se trata el tema de *Recuperación de Pasajes*, estudiando su estado del arte y explicando el fundamento teórico en el que se basa la herramienta de *Recuperación de Pasajes JIRS*, con la cual se realizan los experimentos de los Capítulos 4 y 5.

Capítulo 4, **Recuperación de pasajes en textos legales multilingües**: En este capítulo se expone la dificultad que reviste el proceso de *Recuperación de Pasajes* en textos legales, además, se describen algunas competiciones en las cuales la *Recuperación de Pasajes* es un componente importante para la solución de las mismas. Por último se muestra el trabajo desarrollado y los resultados de la participación en una de estas competencias.

Capítulo 5, **Recuperación de pasajes en patentes multilingües**: Capítulo que describe la dificultad de la *Recuperación de Pasajes* en patentes, también se describen distintas competencias organizadas a nivel mundial sobre el tema. Por último se muestra el trabajo

[12]La colección de documentos *EPO* será estudiada extensamente en la Sección 5.3.1

desarrollado y los resultados de la participación en una de estas competencias.

Capítulo 6, **Conclusiones y trabajo futuro**: En el último capítulo, se analizan los experimentos realizados y los resultados obtenidos para concluir sobre las causas de éxito o posibles errores cometidos a lo largo de las experimentaciones; además, se describen mejoras potenciales en los sistemas que podrían ser estudiadas con profundidad en una investigación doctoral.

Capítulo 2

Procesamiento de textos legales y patentes

El procesamiento de textos legales es un área que presenta grandes dificultades debido a los variados y complejos retos que supone su investigación. El procesamiento de documentos de carácter legal, tales como: patentes, contratos, documentos de ley, etc. es una tarea de creciente importancia en la actualidad, gracias a la necesidad de extraer de ellos datos de relevancia para un fin determinado [19]. La gran cantidad de datos disponibles en el ámbito universitario y empresarial, constituyen una tarea compleja que necesita ser tratada con técnicas adecuadas, para obtener resultados satisfactorios a las necesidades de cada usuario. Según la experimentación realizada en el marco de la presente tesis, se puede concluir que existen tres grandes retos que se presentan en el procesamiento de documentos legales:

1. Cantidad de información: generalmente el volumen de información a procesar en el ámbito legal es elevada, en la Sección 5.3.1 se estudia la colección de documentos *EPO* en la que se aprecia dicha característica;

2. Multilingüismo: una colección de documentos puede presentarse en varios idiomas, lo cual supone un grado de complejidad elevado;

3. Lenguajes y sintaxis propios: los documentos de carácter legal presentan un "lenguaje propio", con términos específicos usados frecuentemente en este tipo de documentos, además, la formulación de los párrafos (secciones) contenidos en los documentos tienen una forma específica, la cual se repite a lo largo de la colección. Supongamos por ejemplo una colección de patentes: en dichos documentos es posible encontrar generalmente, secciones que se refieren al nombre de la patente, o a la descripción de la misma, o al nombre del inventor, etc. Por otra parte la presencia de anáforas, de uso común en el *Lenguaje Natural* y en los documentos de carácter legal, eleva el nivel de complejidad para el análisis de los textos [40]. Para ilustrar el concepto de anáfora supongamos, que en una competencia clásica de *RI*, donde se desea encontrar una respuesta a una

pregunta formulada por un usuario[1], se presenta el siguiente caso:

- Pregunta del usuario: *¿En qué país nació Donald Trump?*

Ahora considere que en un conjunto de documentos se encuentran los dos siguientes pasajes:

- Pasaje 1: *... por tal motivo Donald Trump estuvo de visita en el país en el que nació Mandela, donde ...*
- Pasaje 2: *... El magnate de origen estadounidense visitó el pasado lunes...*

Un sistema clásico de *RI* no estaría calificado para resolver el problema de *BR* adecuadamente, ya que retornaría el primer pasaje como el pasaje más probable en el que se encuentra la respuesta, debido a que en él se encuentran casi la totalidad de los términos de la pregunta. Pero un análisis más profundo, determina la presencia de anáforas con respecto a la pregunta en el segundo pasaje, indicando que en éste pasaje se encuentra con mayor probabilidad la respuesta.

En los textos legales las anáforas son de uso común, normalmente al principio del texto son definidas las partes legales y posteriormente se hace referencia a ellas a lo largo del mismo. En la Figura 2.1 se aprecia un ejemplo real, en el cual se establece un contrato de arrendamiento; en dicha gráfica se aprecia que el señor X y la señora Y son referenciados a lo largo del documento como *"el arrendador/a"*.

Los retos presentes en el procesamiento de documentos legales y patentes anteriormente mencionados, requieren tradicionalmente para su procesamiento, el uso de algoritmos eficientes para obtener resultados satisfactorios.

2.1. Procesamiento de textos legales

El ámbito legal fue uno de los primeros en aplicar las técnicas de *Recuperación de Información* a inicios de los años 60. Consecuencia evidente si se parte de la premisa de que el derecho es una disciplina basada en textos, de hecho es posible que exista una mayor cantidad de textos relativos al derecho que a cualquier otro dominio [8]. La búsqueda de información constituye una parte sustancial del tiempo de un profesional del derecho [48]: el desarrollo de herramientas que faciliten dicha labor demanda un alto interés del mercado global debido a los posibles ahorros de dinero y tiempo invertidos en la tarea.

La *Recuperación de Información* en el ámbito legal se fundamenta en dos métodos:

[1]Competencias conocidas como: competencia de *Búsqueda de Respuesta* ó competencias *BR*


CONTRATO DE ARRENDAMIENTO

CONTRATO DE ARRENDAMIENTO DE VIVIENDA

Contrato Nº

REUNIDOS

DE UNA PARTE Y COMO ARRENDADOR/A:

D. /Dña. X y D./Dña. Y mayor de edad, de nacionalidad ...../vecino de (población) ..... con domicilio en ..... C.P.: ..... provincia de ..... y N.I.F./pasaporte (tachar lo que no proceda) nº .....

Sociedad: ..... con domicilio social en ..... C.P. ..... localidad ..... provincia de ..... C.I.F. nº..... constituida según escritura pública de fecha __ / __ / 20__ otorgada en (localidad) ..... ante el notario D. .....

En representación de ..... Interviene D./Dña ..... mayor de edad, de nacionalidad ..... vecino de (población) ..... con domicilio en ..... C.P. ..... provincia de ..... y N.I.F./pasaporte (tachar lo que no proceda) nº ..... según escritura de poder de fecha __ / __ / 20__ otorgada en ..... ante el notario D. .....

CLÁUSULAS

CLÁUSULAS

PRIMERA.- OBJETO: El arrendador/a cede en arrendamiento al arrendatario/a la vivienda descrita anteriormente, que éste último declara recibir en buen estado de conservación y de habitabilidad. La vivienda arrendada lo es para ser destinada única y exclusivamente a satisfacer la necesidad de vivienda permanente del arrendatario/a.

Figura 2.1: Ejemplo de anáforas en textos legales: contrato de arrendamiento

1. Métodos basados en la ingeniería del conocimiento: enfoque basado en *Inteligencia Artificial* (*IA*) y el razonamiento por medio de casos de estudio.

   Este tipo de métodos generalmente hace uso de ontologías para el procesamiento de los textos. Dicha necesidad abre las puertas de la investigación para la generación de las ontologías. Este es el caso de Lenci et al. [39], el cual pretende realizar la extracción automática de conocimiento ontológico de textos legislativos en el ámbito medioambiental, por medio del uso de máquinas de aprendizaje de lenguajes y análisis estadísticos de textos. Un trabajo parecido es desarrollado por Saias y Quaresma [73], los cuales buscan la creación automática de ontologías a partir de documentos legales, a través del uso de *reconocedores de unidades gramaticales* y *reconocedores de entidades nombradas*, al igual que análisis sintácticos y semánticos de los textos.

   Por otra parte las investigaciones realizadas por Moens et al. [48] y por Mochales et al. [57], centran sus estudios en la detección de argumentos para la clasificación de textos legales, usando clasificadores entrenados por medio de argumentos anotados. La finalidad de las investigaciones es reconocer automáticamente la estructura y los argumentos en un texto jurídico, y la clasificación de uno o varios argumentos en función de su tipo.

   Otros estudios indican que el carácter multilingüe que reviste el ámbito jurídico, plantea la necesidad de contar con "diccionarios" especializados, que permitan la interpretación de los términos jurídicos para ser aplicados por cualquier usuario a nivel mundial; este es el objetivo planteado por la investigación de Ajani et al. [3], donde la creación de

dichos "diccionarios" es guiada por medio de ontologías simples.

2. Métodos basados en *Procesamiento del Lenguaje Natural*: generalmente este tipo de enfoque hace uso de herramientas de base estadística. Sistemas complejos como el desarrollado por Engers et al. [81], el cual pretende hacer una traducción de textos jurídicos en un modelo formal que permita un rápido y fácil entendimiento por parte de un usuario. El sistema propuesto hace uso de *reconocedores de unidades gramaticales*, diccionarios (*lexicones*), gramáticas, entre otras herramientas para generar los modelos representativos de los textos analizados.

Es importante destacar que hoy en día los métodos basados en la ingeniería del conocimiento, en esta determinada tarea, han mostrado mejores resultados que los basados en *PLN*. La utilización de ontologías para el procesamiento de textos legales permite a los sistemas una "más fácil" clasificación de los textos y al mismo tiempo sirven de guía a los usuarios para comprender los resultados del procesamiento [47]. No obstante el potencial de los métodos basados en *Procesamiento de Lenguaje Natural* es prometedor y necesita ser estudiado con profundidad para obtener de él los resultados esperados [81]. También es importante anotar que existen estudios que tratan de fusionar lo mejor de ambos métodos para ofrecer sistemas con mejores prestaciones. El trabajo desarrollado por Casellas et al. [13], plantea como reto concebir un sistema que responda a una pregunta realizada por un abogado sobre un caso en específico; para ello, el experimento realizado cuenta con una base de datos de pares pregunta-respuesta; la pregunta formulada por el usuario en *Lenguaje Natural*, es procesada por medio de ontologías, lo cual permite determinar el dominio específico al que hace referencia, reduciendo el conjunto de posibles respuestas. Por otra parte, el sistema hace uso de medidas de distancia semántica entre la pregunta del usuario y cada una de las preguntas almacenadas en la base de datos, para retornar la respuesta asociada a la pregunta de la base de datos de mejor puntuación por distancia.

El impacto de las facilidades de búsqueda de información que suponen las nuevas tecnologías en el área legal, necesita ser tenido en cuenta desde una perspectiva prudente. El acceso a la información de "todos" los casos relevantes, escenario considerado hasta hace poco tiempo como una cita teórica en el estudio de las leyes, supone un riesgo debido a la carencia de herramientas de procesamiento adecuadas para satisfacer las exigencias de los usuarios. La cantidad de información disponible y la diversidad de interpretaciones que se pueden hacer de esta información, conlleva a la recuperación de información de poca relevancia, que en el peor de los casos puede influir en la toma de malas decisiones [47]. Los sistemas de *RI* jurídicos pretenden satisfacer las necesidades de búsqueda de información de los abogados. Las técnicas de *PLN* cumplen un papel importante en el futuro próximo del área, ya que el *Lenguaje Natural* se plantea como el medio de comunicación entre el mundo real (usuarios) y el mundo computacional. Ejemplo de ello son las competencias de *Búsqueda de Respuesta*, en las que se pretende satisfacer las necesidades de información especificas de un usuario, conociendo de antemano que la información requerida se encuentra en una colección de documentos, que debido a su extensión, precisa del uso de herramientas computacionales para

ser procesada.

El procesamiento de textos legales es un tema de actualidad, en el cual investigadores de todo el mundo enfocan sus trabajos para el desarrollo del área. Prueba de ello es la *Conferencia de Recursos y Evaluación Lingüística* (*LREC*[2] por sus siglas en inglés), la cual tiene por objetivos: proporcionar una visión global del estado del arte, explorar nuevos desarrollos de I + D, realizar el seguimiento de las actividades previstas y en curso, evaluar los usos y las necesidades empresariales, entre otros. En el año 2010 el *LREC* organiza una conferencia[3] en la que se tratarán algunos de los avances en la tarea de *Recuperación de Información* en textos legales.

2.2. Procesamiento de patentes

En el Capítulo 1 se ha hecho mención del volumen creciente de información accesible por un usuario en gran cantidad de áreas. Para dar un ejemplo más concreto de esta realidad se examina el área específica de las patentes. En sólo los Estados Unidos existe un conjunto superior a 5 millones de patentes, lo cual bajo un punto de vista de cantidad de información se traduce entre 100 y 200 GB de texto. Una patente puede variar en cantidad de información, desde unos pocos kilobytes a 1.5 megabytes [36]. En una patente puede ser apreciada una gran cantidad de información separada en campos, en algunas de ellas pueden identificarse hasta 50 campos distintos: un gran número de ellos contienen poca información, pero no por esto menos relevante, algunos de los campos son: título de la patente, nombre de los autores, dirección de los autores, número de la solicitud, número de la patente, fecha de aplicación, fecha de expedición, entre otros. Son relativamente pocos los campos que contienen grandes cantidades de texto narrativo (pero que permiten que herramientas de *PLN* puedan encontrar su campo de acción); entre ellos están: resumen, descripción y reclamaciones [36].

La búsqueda de información es un proceso rutinario llevado a cabo tanto por inventores como por funcionarios de oficinas de patentes. Por una parte la búsqueda permite a los inventores conocer el estado del arte (o "prior art") de la materia que estudian, previniendo posibles reinvenciones y por otra parte, permite a los funcionarios detectar posibles casos de plagio de ideas (propiedad intelectual). Dicha búsqueda debido a los costos en tiempo y dinero ha generado la necesidad de automatizar, en la medida de lo posible el proceso, facilitando el acceso a la información. La necesidad del conocimiento de las publicaciones realizadas a nivel mundial, ha llevado a que varios países ofrezcan interfaces web para la búsqueda en sus bases de datos de patentes. Ejemplo de ello es la *Oficina de Marcas y Patentes de los Estados Unidos*[4]. Gran cantidad de investigaciones utilizan diferentes técnicas para la *Recuperación de Información* de bases de datos de patentes. Entre ellas se destacan:

[2]http://www.lrec-conf.org/
[3]http://www.lrec-conf.org/lrec2010/
[4]http://www.uspto.gov/patft/

- Sistemas de *Recuperación de Información* probabilística para la búsqueda y clasificación de patentes, como el propuesto por la investigación desarrollada por Larkey et al. [36];
- Sistemas para minería de datos en patentes, los cuales usan técnicas de correspondencia y análisis de agrupación de patentes, uno de ellos es descrito en el trabajo desarrollado por Marinescu et al. [46];
- Sistemas que integran herramientas de *Procesamiento del Lenguaje Natural*. Particularmente, el trabajo desarrollado por Osborn et al. [56], los cuales aplican dichas técnicas dentro de un sistema de *Recuperación de Información* basado en vectores para la búsqueda de subconjuntos de patentes.

Capítulo 3

Recuperación de pasajes en textos multilingües

La importancia de la búsqueda de información en documentos (textos en formato digital) se ha incrementado exponencialmente desde la invención de la computadora. Desde los años 50s la cantidad de información que permiten almacenar estos dispositivos ha ido incrementando y con ello, el volumen de datos que deben ser procesados para satisfacer las necesidades de los usuarios; esta necesidad da paso al nacimiento del campo de la *Recuperación de Información* [75]. Los sistemas de *RI* son empleados frecuentemente para intentar encontrar los documentos, de una colección, relevantes a las necesidades de un usuario, expresadas por medio de una pregunta o secuencia de términos. El esquema de funcionamiento de un sistema de *Recuperación de Información* puede ser apreciado en la Figura 3.1. Un sistema de *Recuperación de Pasajes* (*RP*) se diferencia de un sistema de *RI* en que no obtiene documentos completos sino porciones de texto (pasajes) relevantes a la consulta del usuario. En la Figura 3.2 puede ser apreciado el esquema funcional de un sistema de *Recuperación de Pasajes*.

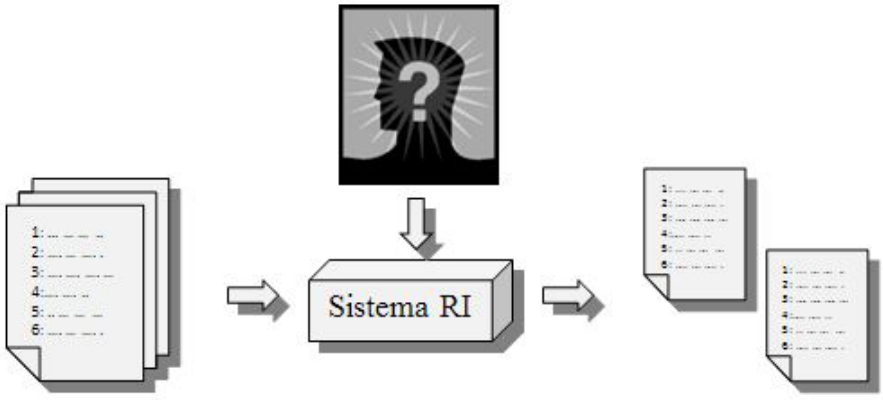

Figura 3.1: Esquema funcional de un sistema de *Recuperación de Información*

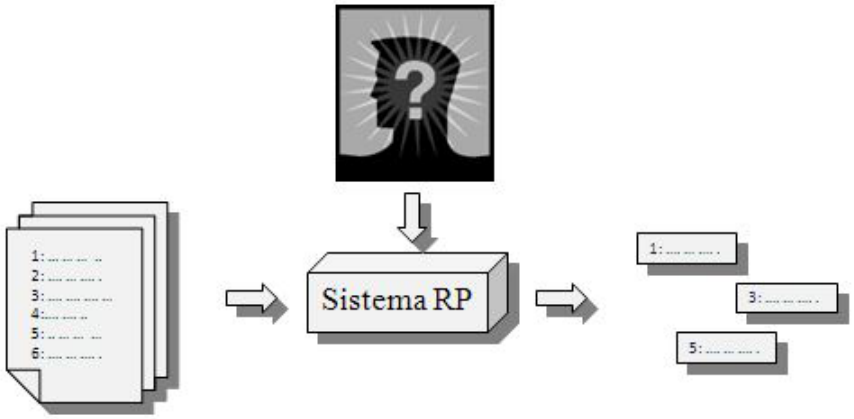

Figura 3.2: Esquema funcional de un sistema de *Recuperación de Pasajes*

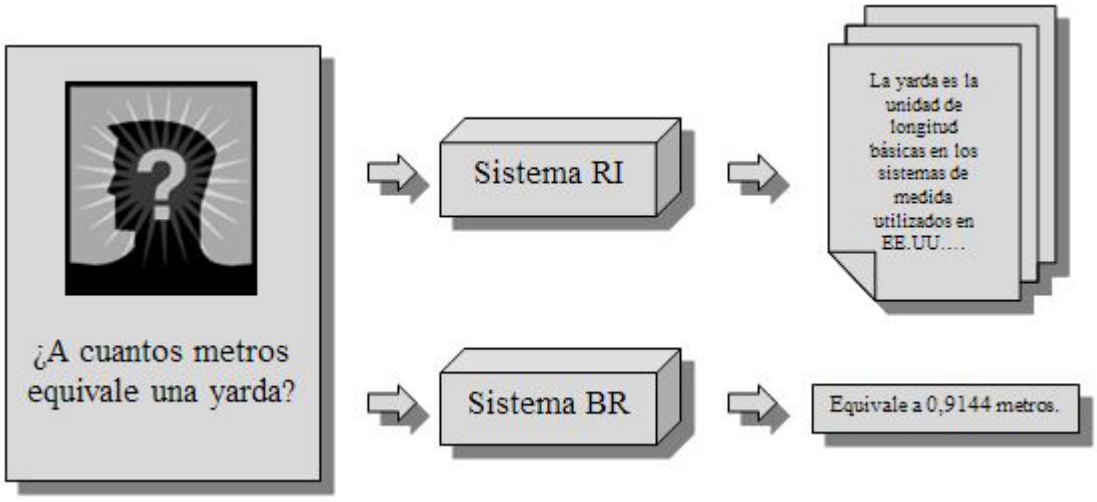

Figura 3.3: Comparación funcional entre Sistemas de *Recuperación de Información* y Sistemas de *Búsqueda de Respuesta*

Generalmente los sistemas de *Búsqueda de Respuesta* (*BR*) son adaptaciones de sistemas de *RI* que pretenden profundizar sobre las necesidades de los usuarios [26]. Un sistema de *BR* responde directamente a las necesidades de los usuarios, ofreciéndoles la menor cantidad de información posible que satisfaga sus necesidades. En la Figura 3.3 se aprecia un ejemplo que evidencia la diferencia entre sistemas de *RI* y sistemas de *BR*.

La implementación de sistemas de *BR* precisa, tradicionalmente, el uso de sistemas de *RP*, con los cuales se obtienen pasajes en los que existe gran posibilidad de encontrar la respuesta [52, 5, 49, 72, 6, 11]. Por lo tanto el sistema de *RP* es un componente de gran importancia para un sistema de *BR*, ya que es un filtro que permite la reducción del espacio de búsqueda de la respuesta de una colección de documentos, a un conjunto de pasajes al-

tamente relacionados con la pregunta realizada por el usuario; en la Sección 4.1 se dará una descripción más detallada de los sistemas de *BR*.

Aproximaciones desde el punto de vista del *Procesamiento del Lenguaje Natural*, se han implementado con el ánimo de obtener mejores resultados [41, 21]; pero dichas aproximaciones son de difícil adaptación a otros lenguajes o tareas multilingües. Como se verá en los resultados de los experimentos descritos en el Capťulo 4, es de gran interés usar enfoques independientes del lenguaje que permitan abordar tareas de carácter multilingüe, de alta demanda en competiciones internacionales.

3.1. Estado del arte

Los sistemas de *RP* empleados en sistemas de *BR* deben cumplir, según Gomez 2004 [26] con:

- Proporcionar pasajes relevantes para la mayor cantidad de preguntas formuladas;
- Proporcionar el mayor número de pasajes relevantes para cada pregunta.

Para cumplir con estos objetivos, se han propuesto diferentes métodos de similitud entre pasaje y pregunta, entre ellos se destacan:

- Métodos donde el cálculo de la similitud entre pasaje y pregunta depende del solapamiento entre sus términos: a mayor solapamiento, mayor similitud. Supongamos por ejemplo la pregunta:
  - *q: ¿Cuál es la velocidad máxima en las autovías alemanas?*

  Y supongamos el pasaje:
  - *p: La velocidad máxima en las autovías alemanas es de 160 km/h.*

  En el ejemplo presentado se aprecia un solapamiento de 7 términos entre pregunta y respuesta; considerando que la pregunta está compuesta por 9 términos, se puede concluir que existe una similitud elevada entre ambos textos.
- Métodos donde la similitud se basa en la densidad de términos de la pregunta y el pasaje. Para ejemplificar el concepto,consideremos la siguiente pregunta:
  - *q: ¿Cuál es la capital de Estados Unidos?*

  Y consideremos los pasajes:

- p_1: *La capital de Estados Unidos es Washington.*
- p_2: *El presidente de la compañía salió de la capital del país, para cerrar la negociación en Estados Unidos.*

En el ejemplo se aprecia como la mayoría de términos de q aparecen "concentrados" en p_1 indicando una alta similitud entre q y p_1, mientras que los términos de q aparecen "dispersos" en p_2 indicando una baja similitud entre ellos.

Estudios de comparación entre las dos aproximaciones mencionadas anteriormente, demuestran que los sistemas basados en densidad de términos presentan un mejor desempeño [78]. Otras aproximaciones destacadas, que no son fácilmente clasificadas en alguno de los dos enfoques mencionados, pero que presentan resultados interesantes son:

- Sistemas basados en densidad de términos, pero que tienen en cuenta el orden de aparición de los términos [2];
- Sistemas basados en el solapamiento de términos con reformulación de la pregunta [76];
- Sistemas que evalúan la similitud entre pregunta y pasaje a nivel sintáctico [77];
- Sistemas que transforman pregunta y pasaje a una representación semántica para determinar su similitud [29];
- Sistemas que implementan patrones léxico-sintácticos para determinar pasajes relevantes [17].

En la siguiente sección se presenta el sistema de *RP* llamado *JIRS*, con el cual se obtienen excelentes resultados en tareas de *BR* [16, 10], y en el cual se fundamentan los experimentos presentados en los Capítulos 4 y 5.

3.2. Sistema *JIRS*

El *Sistema de Recuperación de Información Java*, o por sus siglas en inglés *JIRS*[1] *(Java Information Retrieval System)*, es un sistema de recuperación de pasajes orientado a la búsqueda de respuesta. *JIRS* busca ser una alternativa de mejores prestaciones, comparada con los tradicionales modelos de *Búsqueda de Respuesta*, los cuales basan su funcionamiento generalmente, en análisis estadísticos como el *BM25* [69] o la frecuencia de términos como *TF-IDF* [74], para poder a través de un método heurístico, obtener pasajes relevantes de una colección de documentos susceptibles de ser una respuesta.

[1] http://sourceforge.net/projects/jirs/

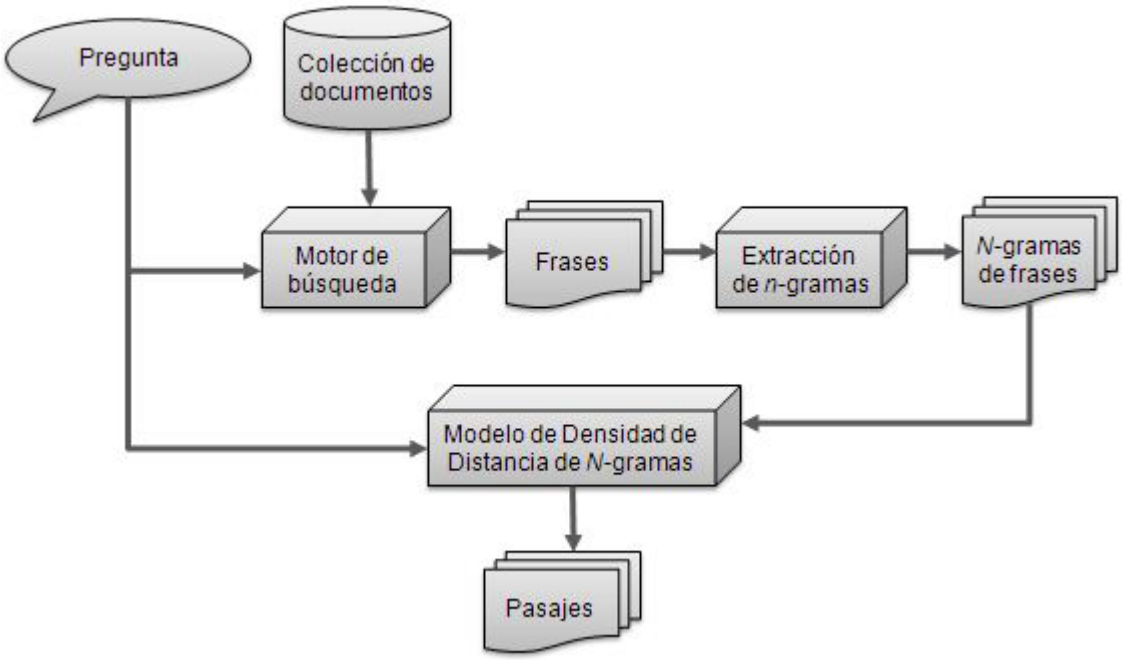

Figura 3.4: Arquitectura de *JIRS*

La hipótesis en la cual se basa el funcionamiento de *JIRS*, supone que en una colección de documentos lo suficientemente grande, es posible encontrar un respuesta a una pregunta formulada, ya que será posible encontrar pasajes cuya sintaxis sea muy parecida a la forma en la que se expresa la pregunta. Para ello *JIRS* analiza estructuralmente, tanto los pasajes como la pregunta y establece una medida de similitud entre ambos.

En la Figura 3.4 se pueden apreciar los principales componentes de la arquitectura con la que se implementa *JIRS*, donde:

- Pregunta: formulada por un usuario en *Lenguaje Natural*, por ejemplo:
  - *¿En qué año se firmó el tratado de Lisboa?*
  - *¿Cuál es el nombre del presidente de Brasil?*
  - *¿Dónde se jugará la copa del mundo del 2014?*
- Colección de documentos: debe estar compuesta por una cantidad de información suficientemente grande. Según la hipótesis en la que se fundamenta *JIRS*, si la cantidad de información es escasa, la posibilidad de que en dicha colección de datos se encuentre una respuesta satisfactoria a la pregunta formulada, es baja.
- Motor de búsqueda: el motor de búsqueda es un sistema clásico de recuperación de información, el cual retorna las frases de la colección de datos en las cuales aparece

alguno de los términos que componen la pregunta. Con el motor de búsqueda se pretende reducir la cantidad de información que tendrá que ser analizada por módulos posteriores.

- Extracción de n-gramas: la extracción de n-gramas "divide" cada uno de los pasajes devueltos por el *motor de búsqueda* en todos los posibles n-gramas que los constituyen. Un ejemplo de la aplicación de la función de *extracción de n-gramas*, de la frase *"El tratado de Lisboa se firmó el 13 de Diciembre de 2007"* se aprecia en la Tabla 3.1.

- Modelo de Densidad de Distancias de N-gramas: *el modelo de densidad de distancia de n-gramas*, es el encargado de evaluar la similitud entre la pregunta y cada uno de los pasajes devueltos por el motor de búsqueda.

  La medida de similitud es definida en la Ecuación 3.1. Dicha ecuación determina el "peso[2]" de un pasaje y de la pregunta, para establecer la similitud entre ambos por medio de una relación. A medida que la relación se aproxime a 1 indicará que la similitud entre pasaje y pregunta es alta o por el contrario si se acerca a 0 indicará que la similitud entre pasaje y pregunta es baja.

$$Sim(p,q) = \frac{W_p}{W_q} \tag{3.1}$$

  Donde:

  - El "peso" asociado a la pregunta (W_q), estará determinado por la Ecuación 3.2:

$$W_q = \sum_{i=1}^{n} w_i \tag{3.2}$$

    Siendo n es el numero de términos de la pregunta y w_i el "peso" asociado a cada término de la misma, el cual es calculado por la relevancia de cada termino en la colección de datos por medio de la Ecuación 3.3:

$$w_i = 1 - \frac{ln(n_i)}{1 + ln(N)} \tag{3.3}$$

    En la cual n_i es el número de pasajes en el que aparece el término analizado t_i. N es el número total de pasajes en la colección de documentos, con lo cual los términos que aparecen con menor frecuencia en la colección de datos tienen asociado un "peso" alto, mientras que los términos que aparecen frecuentemente a lo largo de la colección de datos tienen asociado un "peso" bajo, debido a que estos términos no ofrecen información importante, para determinar si un pasaje ofrece una respuesta satisfactoria a una pregunta determinada. Supongamos que en una colección de datos compuesta por 1.000 documentos, el término "*Lisboa*" aparece

Tabla 3.1: Ejemplo de la función *Extracción de n-gramas*

Frase	El tratado de Lisboa se firmó el 13 de Diciembre de 2007
1-gramas	El
	tratado
	de
	lisboa
	se
	firmó
	el
	13
	de
	diciembre
	de
	2007
2-gramas	El tratado
	tratado de
	de Lisboa
	Lisboa se
	se firmó
	firmó el
	el 13
	13 de
	de Diciembre
	Diciembre de
	de 2007
...	...
4-gramas	El tratado de Lisboa
	tratado de Lisboa se
	de Lisboa se firmó
	Lisboa se firmó el
	se firmó el 13
	firmó el 13 de
	el 13 de Diciembre
	13 de Diciembre de
	de Diciembre de 2007
...	...
11-gramas	El tratado de Lisboa se firmó el 13 de Diciembre de
	tratado de Lisboa se firmó el 13 de Diciembre de 2007
12-gramas	El tratado de Lisboa se firmó el 13 de Diciembre de 2007

Tabla 3.2: Ejemplo de asignación de pesos que emplea *JIRS*. Término: término evaluado, n_i: número de pasajes en el que aparece el término, N: número total de pasajes en la colección de documentos, w_i: peso asociado al término

Término	n_i	N	w_i
Lisboa	250	1.000	0,30
El	950	1.000	0,13

en 250 de ellos y el término "*el*" aparece en 950 documentos. Los pesos asociados a cada termino según la Ecuación 3.3 pueden ser apreciados en la Tabla 3.2.

- El "peso" asociado a cada pasaje (W_p) está determinado por la Ecuación 3.4:

$$W_p = \sum_{\forall x \in Q} h(x, P) \frac{1}{d(x, x_{max})} \tag{3.4}$$

Donde P es el conjunto de n-gramas de mayor "peso" de un pasaje p, cuyos términos aparecen en la pregunta q, teniendo en cuenta que en el conjunto P no puede haber más de un n-grama que comparta algún término; Q es el conjunto de j-gramas que son generados a partir de la pregunta q. Para ilustrar esto se expone el siguiente ejemplo; supongamos que se formule la pregunta:

  - *¿En qué año se firmó el tratado de Lisboa?*

Y supongamos que el pasaje recuperado sea:

  - *El tratado de Lisboa se firmó el 13 de Diciembre de 2007*

El conjunto P del pasaje está compuesto por los n-gramas: *"El tratado de Lisboa"* y *"se firmó"*.

La función $h(x, P)$ determina el peso del j-grama (x) y se define como se muestra en la Ecuación 3.5:

$$h(x, P) = \begin{cases} \sum_{k=1}^{j} w_i & \text{if } x \in P \\ 0 & \text{en otro caso} \end{cases} \tag{3.5}$$

El otro término que determina el "peso" del pasaje es el factor de distancia: $\frac{1}{d(x, x_{max})}$, el cual disminuye el peso de los n-gramas x, a medida que más términos los separan del n-grama de mayor peso x_{max}. La función $d(x, x_{max})$ se define de la siguiente forma:

$$d(x, x_{max}) = 1 + k \cdot\ ln(1 + L) \tag{3.6}$$

[2]Valoración numérica asociada a un pasaje o a una pregunta

Tabla 3.3: Ejemplo del concepto de distancia entre n-gramas de *JIRS*, L: Cantidad de palabras de separación entre los n-gramas, $\frac{1}{d(x,x_{max})}$: Factor de distancia

Pasaje	L	$\frac{1}{d(x,x_{max})}$
El tratado de Lisboa se firmó el 13 de Diciembre de 2007	0	1
El tratado de Lisboa no entró en vigor de inmediato cuando se firmó	2	0.75
El tratado de Lisboa no entró en vigor de inmediato cuando se firmó	7	0.62

En la Ecuación 3.6, el factor k determina la importancia que tiene la distancia en la valoración de la similitud. A medida que el valor de k sea mayor, mayor será la relevancia de la distancia sobre el "peso" de un pasaje. En extensos experimentos llevados a cabo por Gómez et al. [26], se demuestra que el valor de k en el cual el sistema presenta mejores resultados esta en el rango de 0.2 a 0.4. El término L hace referencia a la cantidad de términos que separan a x de x_{max}. Para ejemplificar el concepto de distancia que se aplica en $JIRS$ supongamos que se recuperan los siguinetes pasajes:

- p_1: *El tratado de Lisboa se firmó el 13 de Diciembre de 2007*
- p_2: *El tratado de Lisboa no entró en vigor de inmediato cuando se firmó*

El conjunto P de los n-gramas de los pasajes sería:

- x_1: *El tratado de Lisboa*
- x_2: *se firmó*
- x_3: *en*

Donde se asumirá que el n-grama de mayor peso (x_{max}) es el n-grama x_1. El concepto de distancia que aplica $JIRS$ atenuará en menor medida el peso del pasaje donde los n-gramas analizados se encuentren mas cercanos entre si, tal y como se aprecia en la Tabla 3.3.

Estudios realizados por Buscaldi et al. [11] ilustran el mejor desempeño de *JIRS* ante una tarea de *Búsqueda de Respuesta*, respecto a un sistema de *RI* tradicional como *Lucene*[3]. En la Figura 3.5 se muestra como *JIRS* obtiene una mejor cobertura[4]. De ella se puede interpretar que si ambos sistemas retornan un total de 20 pasajes, existe un posibilidad cercana al 90 % de que en los pasajes devueltos por *JIRS* se encuentre la respuesta a la pregunta, mientras que *Lucene* alcanza un 84 %. Como ya se ha mencionado, JIRS hace uso de un sistema de *RI* tradicional, el cual retorna pasajes en los cuales aparece alguno de los términos incluidos en la pregunta. En la Figura 3.6 se puede apreciar como *JIRS* mejora la calificación de los pasajes

[3]Sistema basado en modelos de espacios vectorial (http://lucene.apache.org/)

[4]Número de pasajes recuperados correctamente respecto al número total de pasajes en la colección de documentos

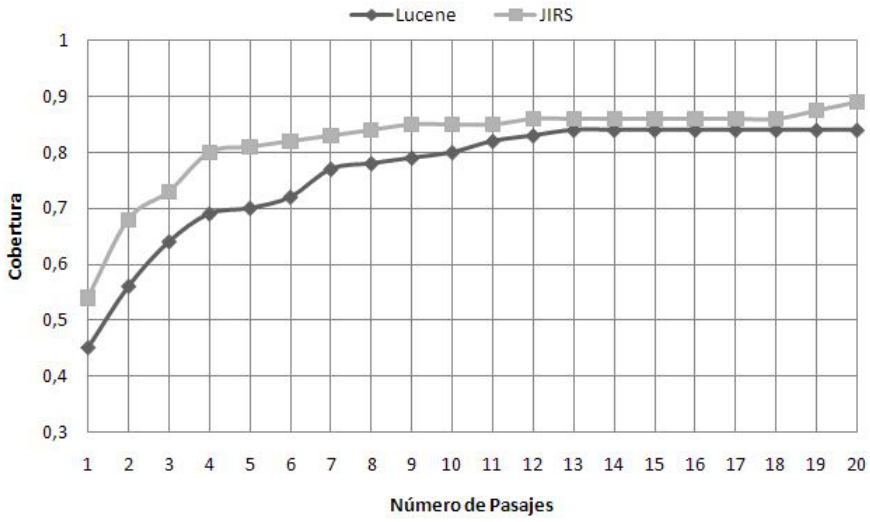

Figura 3.5: Comparación de cobertura de respuestas entre *JIRS* y *Lucene* [11]

devueltos por el motor de búsqueda de *Yahoo*[5], para la tarea de búsqueda de respuesta del *CLEF 2005* [27]. *JIRS* excede en un 22 % la cobertura obtenida por el sistema de *RI* de *Yahoo*, sobre una muestra de 20 pasajes devueltos por cada sistema. En otras palabras, *JIRS* permite mejorar el ranking de los pasajes relevantes devueltos por *Yahoo*.

3.2.1. Ejemplo práctico de funcionamiento del sistema *JIRS*

A continuación se aprecia un ejemplo práctico de la teoría vista en la Sección 3.2.

Suponga que desea buscar en una colección de datos de una agencia de noticias, la respuesta a la pregunta:

- *q: ¿En qué año se firmó el tratado de Lisboa?*

Ahora suponga que el *motor de búsuqeda* retorna los siguientes pasajes, que serán evaluados por el *modelo de densidad de distancia de n-gramas*:

- p_1*: El tratado de Lisboa se firmó el 13 de Diciembre de 2007*
- p_2*: El tratado de Lisboa no entró en vigor de inmediato cuando se firmó*

Una vez procesada la colección de datos se obtienen los "pesos" de todos los terminos. En la Tabla 3.4 se aprecian los valores del procesamiento[6].

[5] http://www.yahoo.com/

[6] Valores supuestos para su uso en el ejemplo.

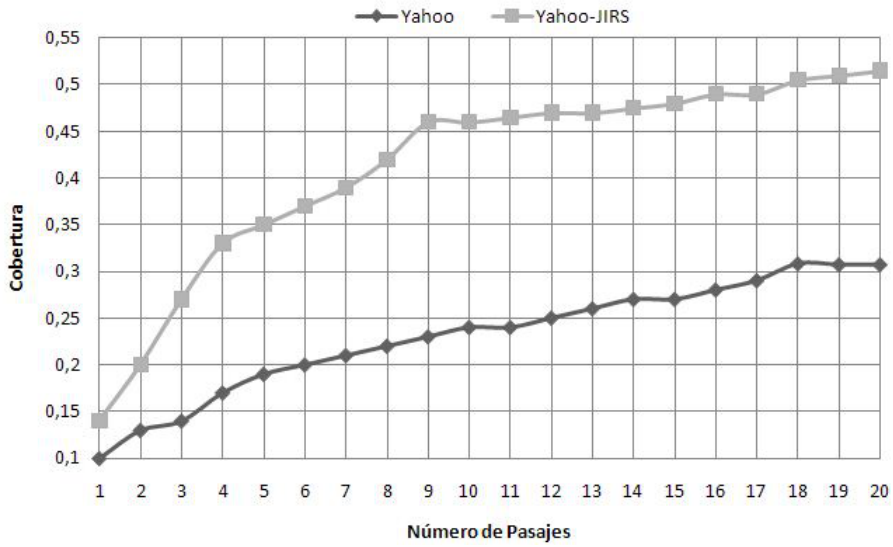

Figura 3.6: Comparación de cobertura de respuestas entre *JIRS* y *Yahoo-JIRS* [11]

Tabla 3.4: Pesos de términos para efectos de ejemplo. Término: término evaluado, w: peso asociado al término

Término	w
en	0.1
que	0.1
año	0.2
se	0.1
firmó	0.5
el	0.1
tratado	0.5
de	0.1
lisboa	0.6
13	0.2
diciembre	0.1
2007	0.2
no	0.1
entró	0.1
vigor	0.3
inmediato	0.2
cuando	0.1

Tabla 3.5: Cálculo del factor de distacia para el pasaje: ‘‘El tratado de Lisboa se firmó el 13 de Diciembre de 2007”

Pasaje	x	$\frac{1}{d(x,x_{max})}$
El tratado de Lisboa se firmó el 13 de Diciembre de 2007	x_1	1
El tratado de Lisboa se firmó el 13 de Diciembre de 2007	x_2	1

Tabla 3.6: Cálculo del factor de distacia para el pasaje: *El tratado de Lisboa no entró en vigor de inmediato cuando se firmó*

Pasaje	x	$\frac{1}{d(x,x_{max})}$
El tratado de Lisboa no entró en vigor de inmediato cuando se firmó	x_1	1
El tratado de Lisboa no entró en vigor de inmediato cuando se firmó	x_2	0.62
El tratado de Lisboa no entró en vigor de inmediato cuando se firmó	x_3	0.75

Según los datos de la Tabla 3.4 y según la Ecuación 3.2, el “peso” asociado a la pregunta es: $W_q = 2.1$

El conjunto P_1 para p_1 sería:

- x_1: *El tratado de Lisboa*
- x_2: *se firmó*

Mientras que el conjunto P_2 para p_2 sería:

- x_1: *El tratado de Lisboa*
- x_2: *se firmó*
- x_3: *en*

Los pesos de los pasajes de los conjuntos P_1 y P_2, según los datos del Tabla 3.4, son:

- $W_{x_1} = 1.3$
- $W_{x_2} = 0.4$
- $W_{x_3} = 0.1$

Los factores de distancia para p_1 y p_2 son apreciados en las Tablas 3.5 y 3.6 respectivamente, teniendo en cuenta que x_{max} según los pesos de los pasajes es x_1.

En el Tabla 3.3 se pueden apreciar los factores de distancia, por lo tanto las similitudes son calculadas de la siguiente forma:

$$Sim(p_1, q) = \frac{W_{x_1} \cdot \frac{1}{d(x_1,x_1)} + W_{x_2} \cdot \frac{1}{d(x_2,x_1)}}{W_q} = 0.66 \tag{3.7}$$

$$Sim(p_2, q) = \frac{W_{x_1} \cdot \frac{1}{d(x_1,x_1)} + W_{x_2} \cdot \frac{1}{d(x_2,x_1)} + W_{x_3} \cdot \frac{1}{d(x_3,x_1)}}{W_q} = 0.60 \tag{3.8}$$

JIRS asigna mayor valor a la similitud entre p_1 y q, donde efectivamente se encuentra una respuesta satisfactoria a la pregunta formulada. También es interesante reflexionar en el hecho de que gracias al factor de distancia, la similitud entre p_1 y q obtiene un mayor valor, en comparación con el valor asociado a la similitud entre p_2 y q, no obstante p_1 contener menos cantidad de n-gramas incluidos en q.

Capítulo 4

Recuperación de pasajes en textos legales multilingües

La *Recuperación de Pasajes* en textos legales, intenta satisfacer la necesidad de información específica de los usuarios, en dominios de carácter legal. Por ejemplo:

- En el ámbito de las leyes aplicables en los países de la comunidad europea, recopilados en la colección de documentos *JRC-Aquis* estudiada en la Sección 4.3.1;
- A un nivel empresarial, las leyes que rigen su actividad productiva en el código de comercio del país en el que actúan;
- Los documentos legales que hacen parte inherente de la conformación de una empresa, como lo son los documentos de contratación de personal y servicios.

Varios autores exponen la necesidad de contar con sistemas automáticos, que faciliten la búsqueda eficiente de información en el ámbito legal, ya que actualmente este tipo de labor es desarrollado manualmente por un experto, lo cual se traduce en costos elevados para las empresas que lo requieran [81, 48].

Los retos propuestos por la tarea de *Recuperación de Pasajes* en textos legales, generalmente conllevan enfoques que revisten una alta complejidad como se verá a lo largo de este capítulo. Entre los retos más interesantes a los cuales se hace referencia anteriormente se encuentran:

- Anáforas: la presencia de anáforas es una característica que se presenta frecuentemente en textos legales. Por ejemplo, suponga que se desea encontrar pasajes referentes a la *"venta de las acciones de la empresa X"*, puede darse el caso en el que un documento hable de la *empresa X*, pero el pasaje que habla de la venta de las acciones de la

empresa, se refiera a ella como: *"la gran empresa de venta de productos Y"*, por lo que es necesario que los sistemas sean capaces de determinar que *"la empresa X"* y *"la gran empresa de venta de productos Y"* hacen referencia a la misma entidad.

- Temporalidad: la búsqueda de pasajes que revisten características temporales, se presentan contínuamente en cualquier tipo de documentos. Por ejemplo, suponga que un usuario desea saber cuál es el presidente de la *empresa X*. En los documentos en los que se desea realizar la búsqueda, puede encontrarse información de dos personas que han sido presidentes de la misma empresa pero en distintos períodos de tiempo. Debido a esta particularidad es deseable que el sistema sea capaz de inferir a que periodo de tiempo se refiere el usuario.

Las propuestas realizadas en este documento y más específicamente el uso de un enfoque multilingüe que hasta el momento no había sido explorado en competencia [59], muestran cómo con un enfoque simple se pueden obtener excelentes resultados, abriendo un campo interesante de estudio para futuras herramientas de recuperación de pasajes multilingües en textos legales.

El capítulo está conformado de la siguiente forma: en la Sección 4.1 se hace un análisis del estado del arte de la tarea de búsqueda de respuesta; posteriormente en las Secciones 4.2 y 4.3 se hace referencia a las competiciones de búsqueda de respuesta del *CLEF*, competición en la que se evalúan los experimentos estudiados en el desarrollo de esta investigación; un estudio a profundidad de los enfoques propuestos en el desarrollo de la investigación se aprecia en la Sección 4.4; en la Sección 4.5 se describen los experimentos realizados; por último en la Sección 4.6 se hace un análisis de los resultados obtenidos.

4.1. Búsqueda de Respuestas

Los sistemas de *RI* han demostrado ser eficaces en grandes colecciones de documentos para determinar el conjunto de documentos relevantes a un tema puntual, pero en algunos casos es deseable ser un poco más específico y conocer la respuesta exacta a una pregunta. Los sistemas de *BR* intentan encontrar respuestas concretas a preguntas precisas formuladas por un usuario [26].

El concepto de sistemas de *BR* inicia en los años 60 con sistemas que actuaban en un dominio restringido; generalmente, dichos sistemas contaban con una base de datos escrita a mano por los expertos del dominio elegido, como por ejemplo el sistema *Baseball*, el cual responde preguntas de la liga norteamericana de baseball de un año en específico [28]. El avance de la lingüística computacional en las décadas de los años 70 y 80, permitió el desarrollo de importantes proyectos, que pretendían con base en un conjunto de conocimientos, encontrar respuestas a preguntas afines a los conocimientos disponibles, ejemplo de ello es el sistema *Unix Consultant*, sistema que responde a preguntas referentes al sistema *Unix* [84]. Desde

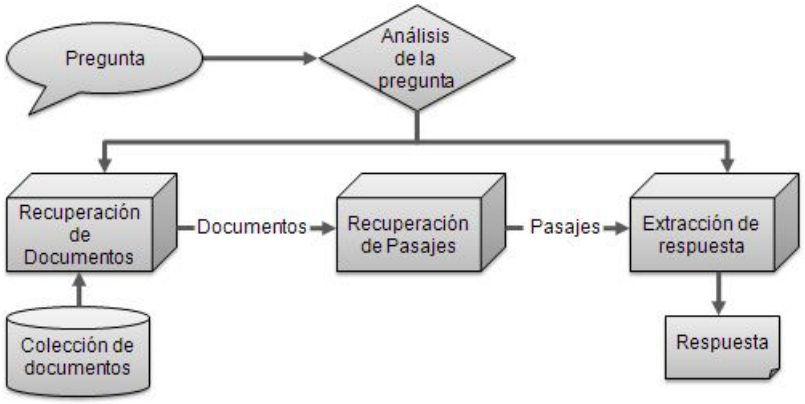

Figura 4.1: Estructura de un Sistema de *Búsqueda de Respuesta*

finales de los 90 hasta hoy, diferentes competiciones a nivel mundial son organizadas con el objetivo de promover la investigación y el desarrollo de sistemas de *Búsqueda de Respuesta*, en los que generalmente se intenta responder a diferentes preguntas, basándose en conjuntos de documentos de distribución libre que tratan una gran diversidad de temas y que a su vez, constituyen la "base de conocimiento" de los sistemas.

En la actualidad los sistemas de *BR* que presentan mejores resultados [59] se basan en la estructura de la Figura 4.1, donde:

- Análisis de la pregunta: es el módulo encargado de determinar el tipo de respuesta esperado. Por ejemplo, para la pregunta: *¿Dónde se encuentra la estatua de la libertad?*, la respuesta esperada se referirá a un lugar;
- Recuperación de documentos: este módulo extrae los documentos de la colección, relevantes a la pregunta;
- Recuperación de pasajes: el objetivo del módulo de *RP* es extraer de un documento relevante, un fragmento en el cual se encuentre con mayor probabilidad la respuesta a la pregunta formulada;
- Extracción de respuesta: analiza los pasajes para determinar la posible respuesta.

La fiabilidad de los sistemas de *BR* actuales, basados en la estructura de la Figura 4.1 retornan resultados interesantes, si se permite al sistema indicar no sólo una única respuesta, sino un conjunto de posibles respuestas, las cuales pueden ser analizadas por el usuario con el fin de determinar cuál de ellas responde a sus necesidades. Por ejemplo si el sistema se restringe a sólo una respuesta como salida, se obtienen resultados en competencia alrededor del 50 %, pero si se permite al sistema retornar una lista de 20 posibles respuestas a la pregunta formulada, se pueden obtener resultados cercanos al 90 % [11, 15].

Los tres grandes foros que promueven la investigación en tecnologías de acceso a la información: NTCIR, TREC y CLEF, referenciados en el Capítulo 1, tradicionalmente contemplan la realización de competencias en el ámbito de *Búsqueda de Respuesta*. En particular, en las Secciones 4.2 y 4.3 se estudiará a profundidad la competencia de *Búsqueda de Respuesta* del foro *CLEF*.

4.2. Competiciones de Búsqueda de Respuesta del *CLEF*

El foro Europeo *CLEF* promueve la investigación y el desarrollo de acceso a información multilingüe con dos objetivos definidos:

1. Desarrollo de la infraestructura para la comprobación, ajuste y evaluación de sistemas de *Recuperación de Información* que operan en lenguas europeas;

2. Creación de bancos de prueba de datos reutilizables, que puedan ser empleados para el desarrollo de sistemas con fines de evaluación comparativa.

Son diversas las competencias organizadas anualmente por el foro, entre ellas se encuentra la competencias de *Búsqueda de Respuesta*. Dicha competencia se ha realizado continuamente durante los últimos 8 años, facilitando un espacio en el cual los diferentes equipos pueden poner a prueba sus enfoques para dar soluciones a las problemáticas planteadas y dar cabida a la comparación, por medio de los resultados de la eficacia de los sistemas participantes.

Una prueba piloto de la competencia se realizó en el año 2003 [44]. Dos tareas fueron propuestas en dicha edición: tarea monolingüe y bilingüe. La tarea monolingüe consistía en encontrar la respuesta en el idioma en el que estaban formuladas las preguntas, mientras que la tarea bilingüe precisaba encontrar las respuestas en un idioma distinto al que se formulaban la preguntas. En ambas tareas un total de 200 preguntas fueron formuladas y los sistemas debían devolver 3 respuestas exactas (o cadenas de caracteres), de un conjunto de documentos proporcionado en el cual se encontraban las respuestas. En total 8 equipos participaron en las tareas propuestas; sistemas complejos y elaborados presentaban enfoques basados en técnicas clásicas de *RI* tales como: *tokenización* y *etiquetado-pos*, *buscadores booleanos* eran empleados para establecer pasajes candidatos a ser soluciones, los cuales eran posteriormente filtrados por módulos de validación de respuestas, basados en *Reconocimiento de Entidades Nombradas* [53].

En 2004 se planteó un ejercicio similar al propuesto en la versión anterior [45], pero con la diferencia de restringir los sistemas a sólo poder retornar una única respuesta a cada pregunta. En total 18 equipos participantes propusieron enfoques que mostraban un mejor desempeño, comparado con los resultados obtenidos en la versión 2003. Entre ellos se destacan enfoques que emplean métodos de *RI* de notable desempeño, como el modelo probabilístico de *OKAPI*

(BM25) [69] o motores de búsqueda como *Lucene* y el uso de técnicas de análisis sintáctico en preguntas y pasajes del conjunto de documentos, para establecer la probabilidad de encontrar las respuestas [54, 61].

La versión 2005 de la tarea de *Búsqueda de Respuesta* del *CLEF* [80], da continuidad a la propuesta realizada por las competencias anteriores, incorporando como novedad preguntas de carácter temporal elevando la dificultad de la competencia. Un total de 24 equipos participaron con enfoques similares a los desarrollados en la pasada edición de la competencia; no obstante, uno de los enfoques presentó una propuesta interesante, la cual realizaba validación vía web de las respuestas [55].

Tres tipos de preguntas nuevas - preguntas de hecho, de definición y de listas - son la novedad en la competencia de *Búsqueda de Respuestas* del *CLEF 2006* [43]. En esta ocasión los sistemas debían retornar tanto la respuesta exacta a la pregunta, como el texto del conjunto de documentos que soporta dicha respuesta. Nuevamente el número de participantes incrementó con respecto a las ediciones pasadas de la competencia, sumando un total de 30 equipos. Los sistemas participantes muestran una elevada complejidad en cuanto a las técnicas utilizadas para su desarrollo. Entre las técnicas que mejores resultados aportaron a la competencia se encuentran: lematización realizada tanto en preguntas, como en el conjunto de documentos, desambiguación gramatical y pesado de términos por medio de un análisis semántico [37].

Para el año 2007, la organización de la competencia decidió no hacer cambios sobre las reglas de la competencia, con el ánimo de refinar los sistemas participantes en la edición anterior [23]. En esta ocasión participaron 22 equipos y se observó el incremento de la complejidad en los sistemas. Técnicas como analizadores sintácticos y semánticos, sistemas clásicos de *RI*, *Reconocimiento de Entidades Nombradas*, uso de Ontologías, módulos de desambiguación, entre otras, fueron implementadas en el sistema de mejor desempeño de dicho año [9].

En 2008 se utilizó *Wikipedia*[1] como conjunto de documentos para la búsqueda de las respuestas a las preguntas formuladas, las reglas para esta competencia fueron iguales a las del año 2007 [18]. La tendencia de implementación de sistemas complejos [4], usando las mismas técnicas empleadas en las competencias anteriores para la solución de la tarea, fue observada en la competencia que contó con un total de 21 equipos participantes.

Para la edición 2009 de la competencia se introdujeron varios cambios que serán estudiados en la Sección 4.3. Actualmente se está organizando la edición 2010 de la competencia, en la cual además de participar con las técnicas de mejor desempeño en las ediciones pasadas, serán propuestas algunas variaciones tales como el uso de una técnica de ranking de pasajes

[1] http://es.wikipedia.org/

basada en el *modelo probabilístico de OKAPI (BM25)*, y el pre-procesamiento del conjunto de documentos basado en la técnica de *stemming*.

4.3. Competición *ResPubliQA* del *CLEF*

El objetivo de la tarea *ResPubliQA-2009* es encontrar la respuesta a un conjunto de 500 preguntas formuladas en *Lenguaje Natural*; supongamos por ejemplo, que un usuario desea obtener respuestas a preguntas del dominio legal, específicamente de la legislación Europea, lo cual supone que el sistema debe ser capaz de interpretar la pregunta del usuario, para buscar su respuesta en un conjunto de datos de gran tamaño. El conjunto de documentos empleado para la realización de la tarea es una sub-colección de documentos del corpus *JRC-Acquis*, cuyas características son estudiadas en la Sección 4.3.1.

Según los lineamientos de la competencia, la respuesta a cada pregunta formulada debe ser la extracción de un párrafo del conjunto de documentos; en ningún caso debía ser la respuesta exacta. El retornar un párrafo completo en vez de una respuesta exacta permite la comparación entre las aproximaciones de *RI* puras y las tecnologíaías actuales de *BR* [59]. Una de las principales motivaciones para el desarrollo de la tarea propuesta es la búsqueda de soluciones a problemáticas reales en un dominio de usuarios potenciales.

Diferentes tipos de preguntas son formuladas en la competencia, los cuales pueden ser clasificadas en las categorías: *Hecho, Definición, Razón, Propósito* y *Procedimiento* [59].

- Las preguntas de *Hecho* buscan responder a temas como: el nombre de una persona, una locación, la fecha en la que ocurrió un evento, etc. Por ejemplo: *¿Cuándo instó el Consejo de Europa una estrategia para mejorar la interoperabilidad de los trenes?*
- Las preguntas de *Definición* investigan el significado de algo, por ejemplo: *¿Qué significa "consumo inmediato"?*
- Preguntas de *Razón* son preguntas que tienen como objetivo encontrar las razones, motivos o motivaciones para que algo pase, por ejemplo: *¿Por qué razón se recomienda tomar muestras de leche a través de la boca del recipiente?*
- Una pregunta de *Propósito* se refiere a una pregunta que busca el objetivo o la meta de algo, por ejemplo: *¿Cuál es propósito de promocionar el transporte ferroviario?*
- Las preguntas de *Procedimiento* buscan determinar un conjunto de acciones que son aceptadas para hacer algo, por ejemplo: *¿Bajo qué condiciones deben ser reconocidos un documento de conformidad y un certificado de gestión de seguridad?*

La clasificación de preguntas que se presentó anteriormente permite implementar sistemas con enfoques eurísticos [12] para determinar las repuestas. Como se verá en la Sección 4.4, el enfoque que hemos propuesto, basado en el análisis de n-gramas, no tiene en cuenta el típo de pregunta a la que se enfrenta, lo cual permite flexibilidad ante posibles modificaciones de la tarea.

Una característica importante de la tarea propuesta por *ResPubliQA-2009* es el carácter multilingüe de la competencia. Al contar con un conjunto de documentos paralelo, como se verá en la Sección 4.3.1, los equipos pueden escoger participar en tareas monolingües donde preguntas y respuestas son formuladas en el mismo idioma, o tareas multilingües donde preguntas y respuestas se presentan en diferentes idiomas. Dicha característica, perite la comparación de los sistemas independientemente del idioma en el cual se participe. Los idiomas encontrados en la sub-colección de documentos del corpus *JRC-Acquis* son: Búlgaro, Holandés, Inglés, Francés, Alemán, Italiano, Portugués, Rumano y Español.

Los sistemas pueden responder a cada pregunta por medio de dos tipos de respuesta:

1. El párrafo exacto extraído del conjunto de documentos;
2. La palabra NOA, que indica que el sistema prefiere no responder a la pregunta.

Las respuestas tienen efecto directo sobre la calificación como se verá en la Sección 4.3.2.

4.3.1. Colección de documentos *JRC-Aquis*

JRC-Acquis[2] es un conjunto de documentos de libre distribución, que comprende documentos generalmente de carácter jurídico, en los cuales se describen las leyes aplicadas en los estados miembros de la *Unión Europea* (*UE*). Entre los tipos de documentos que se pueden encontrar en *JRC-Aquis* están:

- Principios y objetivos políticos de los tratados de la *UE*;
- La legislación de la *UE*;
- Declaraciones y resoluciones;
- Acuerdos internacionales;
- Actos y objetivos comunes.

Los temas que se tratan en el conjunto de documentos son diversos, entre ellos se destacan: Economía, Salud, Tecnología de la información, Derecho, Agricultura, Alimentación y

[2] http://langtech.jrc.it/JRC-Acquis.html

Tabla 4.1: Estadísticas de la coleeción de documentos *JRC-Acquis*. ND: Número de documentos, NP: Número de Palabras, NH: Número de caracteres, NPP: Número de palabras promedio por documento

Lenguajes	ND	NP	NH	NPP
bg	11.384	16.140.819	104.522.671	1417,85
cs	21.438	22.843.279	148.972.981	1065,55
da	23.624	31.459.627	213.468.135	1331,68
de	23.541	32.059.892	232.748.675	1361,87
el	23.184	36.453.749	239.583.543	1572,37
en	23.545	34.588.383	210.692.059	1469,03
es	23.573	38.926.161	238.016.756	1651,3
et	23.541	24.621.625	192.700.704	1045,9
fi	23.284	24.883.012	212.178.964	1068,67
fr	23.627	39.100.499	234.758.290	1654,91
hu	22.801	28.602.380	213.804.614	1254,44
it	23.472	35.764.670	230.677.013	1523,72
lt	23.379	26.937.773	199.438.258	1152,22
lv	22.906	27.592.514	196.452.051	1204,6
mt	10.545	20.926.909	128.906.748	1984,53
nl	23.564	35.265.161	231.963.539	1496,57
pl	23.478	29.713.003	214.464.026	1265,57
pt	23.505	37.221.668	227.499.418	1583,56
ro	6.573	9.186.947	60.537.301	1397,68
sk	21.943	26.792.637	179.920.434	1221,01
sl	20.642	27.702.305	178.651.767	1342,04
sv	20.243	29.433.037	199.004.401	1453,99
Total	463.792	636.216.050	4.288.962.348	1387,23

Política. La colección es continuamente actualizada y cuenta con textos que datan desde 1950 hasta hoy. Una de las principales características de la colección de datos y una de las que despierta mayor interés en el ámbito universitario, es su carácter multilingüe; contando con alineaciones por párrafos para 231 pares de idiomas. *JRC-Aquis* es una colección de textos paralelos, en 22 idiomas: Búlgaro (bg), Checo (cs), Danés (da), Alemán (de), Griego (el), Inglés (en), Español (es), Estonio (et), Finés (fi), Francés (fr), Húngaro (hu), Italiano (it), Lituano (lt), Letón (lv), Maltés (mt), Holandés (nl), Polaco (pl), Portugués (pt), Rumano (ro), Eslovaco (sk), Esloveno (sl) y Sueco (sv). La publicación de datos por medio de *Joint Research Centre*[3] (*JRC*) cuenta con el esfuerzo de la Comisión Europea[4] con el fin de apoyar el multilingüismo, la diversidad lingüística y la reutilización de la información de la Comisión. En la Tabla 4.1 se aprecian algunas estadísticas de la colección de documentos.

Los documentos de *JRC-Acquis* están codificados en formato XML, de acuerdo a las directrices *TEI*[5]. Un ejemplo de la codificación empleada puede ser apreciada en la Figura

[3]http://ec.europa.eu/dgs/jrc/index.cfm
[4]http://ec.europa.eu/index_en.htm
[5]http://www.tei-c.org/Guidelines/

```
<TEI.2 id=''id'' lang=''lenguaje''>
    <teiHeader lang=''lenguaje'' date.created=''fecha''>
        <fileDesc>
            <titleStmt>
                <title>Título del documento</title>
            </titleStmt>
        </fileDesc>
    </teiHeader>
    <text>
        <body>
            <head n=''1''>Párrafo 1</head>
            <div type=''body''>
                <p n=''2''>Párrafo 2</p>
                <p n=''3''>Párrafo 3</p>
                ...
                <p n=''m''>Párrafo m</p>
            </div>
        </body>
    </text>
</TEI.2>
```

Figura 4.2: Formato XML empleado en la colección de documentos *JRC-Acquis*

4.2.

JCR-Acquis es uno de los conjuntos de documentos paralelos más grandes a nivel mundial, empleado satisfactoriamente en diversas tareas entre las cuales se destacan: entrenamiento de sistemas automáticos, tareas de traducción automática estadística; producción de léxico multilingüe o recursos semánticos como diccionarios u ontologías, entrenamiento de clasificadores de dominio temático multilingüe y competiciones de carácter multilingüe.

4.3.2. Medidas de evaluación

Dos medidas de evaluación han sido utilizadas para determinar el desempeño de los equipos participantes en la competición *ResPubliQA* del año 2009 [59]:

1. *Exactitud*: dadas las características de la competencia, se define la exactitud como la proporción de respuestas correctas, considerando las respuestas candidatas de preguntas a las que se prefiere no contestar, respecto al número de preguntas formuladas. La *Exactitud* es definida de la siguiente manera:

$$Exactitud = \frac{C + NOC\_C}{N} \tag{4.1}$$

   Donde:

   - C: número de respuestas correctas;
   - NOC_C: número de preguntas no contestadas en las cuales la respuesta candidatas son correctas;

- N: número de preguntas.

2. *c@1*: gracias a la posibilidad de no contestar a una pregunta, la medida *c@1* pretende "premiar" a los sistemas capaces de determinar la conveniencia o no, de responder a una pregunta dada la información que han recuperado. La medida c@1 se define en la Ecuación 4.2:

$$c@1 = \frac{1}{N}(C + NOC.\frac{C}{N}) \tag{4.2}$$

Donde:

- C: número de respuestas correctas;
- NOC: número de preguntas no contestadas;
- N: número de preguntas.

Se puede apreciar que las medidas *Exactitud* y $c@1$ son iguales si el número de preguntas no contestadas es igual a cero, pero si un sistema es capaz de no contestar a algunas preguntas de manera acertada, la medida $c@1$ incrementara el valor de la calificación de la participación.

4.4. Sistema de Recuperación de Pasajes para el dominio legal

Para resolver el problema planteado por la competencia *ResPubliQA 2009* expuesto en la Sección 4.3, se han implementado enfoques basados en n-gramas mediante el uso de la herramienta *JIRS* explicada en la Sección 3.2. Con dichos enfoques se pretende implementar una herramienta de baja complejidad, que permita tener resultados competitivos, comparándola con sistemas más elaborados, como por ejemplo el sistema implementado por Rodrigo et al. [71], el cual involucra procesos de n-gramas, *Reconocimiento de Entidades Nombradas*, *Expresiones Temporales* y *Expresiones Numéricas*; o el sistema implementado por Ion et al. [32], el cual emplea procesos de *Reconocimiento de Entidades Nombradas* y *Expresiones Numéricas*. Es importante señalar que al implementar un enfoque únicamente basado en n-gramas, se obtiene un sistema independiente del idioma, lo cual abre las puertas a participar en prácticamente cualquier sub-tarea (véase Sección 4.3), realizando pocos cambios en la configuración del sistema. Dos enfoques se han empleado para participar en *ResPubliQA 2009*: enfoque monolingüe y enfoque multilingüe. Con ello se pretende determinar la incidencia o no del paralelismo idiomático, que provee la sub-colección de documentos del conjunto de documentos *JRC-Acquis* (véase Sub-sección 4.3.1). En las siguientes secciones se explica en detalle cada uno de ellos.

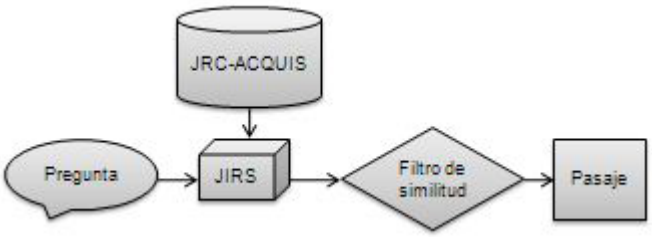

Figura 4.3: Esquema del Sistema de *Recuperación de Pasajes* para la competencia *ResPubliQA 2009*, enfoque monolingüe

4.4.1. Enfoque monolingüe

El enfoque monolingüe apreciado en la Figura 4.3, es un enfoque simple que pretende corroborar la hipótesis en la que se fundamenta *JIRS* (Véase Sección 3.2), la cual propone que en una colección de documentos lo suficientemente grande, es posible encontrar un respuesta a una pregunta formulada. Por lo tanto, con el enfoque propuesto se pretende determinar la aptitud de la herramienta, para la competencia de *Búsqueda de Respuesta* propuesta por *ResPubliQA 2009* en el dominio legal.

El enfoque propuesto calcula la similitud entre la pregunta formulada y cada uno de los pasajes de la colección de documentos en un mismo idioma. Cada pregunta formulada es analizada por el sistema, el cual selecciona como respuesta el pasaje de la colección de documentos que presente una mayor similitud. El *filtro de similitud*, ubicado a la salida del sistema, pretende eliminar los pasajes que aún siendo los mejores calificados por el sistema presentan una "baja" similitud con la pregunta. Supongamos por ejemplo, que entre todos los pasajes de la colección de documentos el mejor calificado obtiene una puntuación de 0,1; según lo explicado en el Sección 3.2, la puntuación varia en el rango [0,1] indicando que a mayor puntuación mayor similitud, por lo tanto una puntuación tan baja indicará que dicho pasaje tiene una gran posibilidad de no satisfacer las necesidades del usuario.

4.4.2. Enfoque multilingüe

Con el enfoque multilingüe ilustrado en la Figura 4.4, se pretende determinar la incidencia del carácter paralelo, en cuanto a idiomas se refiere, que presenta la colección de documentos *JRC-Acquis* en la tarea propuesta por *ResPubliQA 2009*. Los documentos encontrados en *JRC-Acquis* son formulados en un idioma y posteriormente son traducidos al resto de idiomas presentes en la *UE*. La traducción de los documentos permite plantear la siguiente hipótesis: "*Es posible encontrar una mayor similitud entre una pregunta y un pasaje, si se maximiza dicha similitud analizándola en un ámbito multilingüe.*" En la Ecuación 4.3 se aprecia cómo aplicar el anterior concepto a la tarea propuesta. La respuesta p a una pregunta q, estará determinada por la puntuación más alta de la similitud entre la pregunta y cada pasaje evaluado en el idioma x.

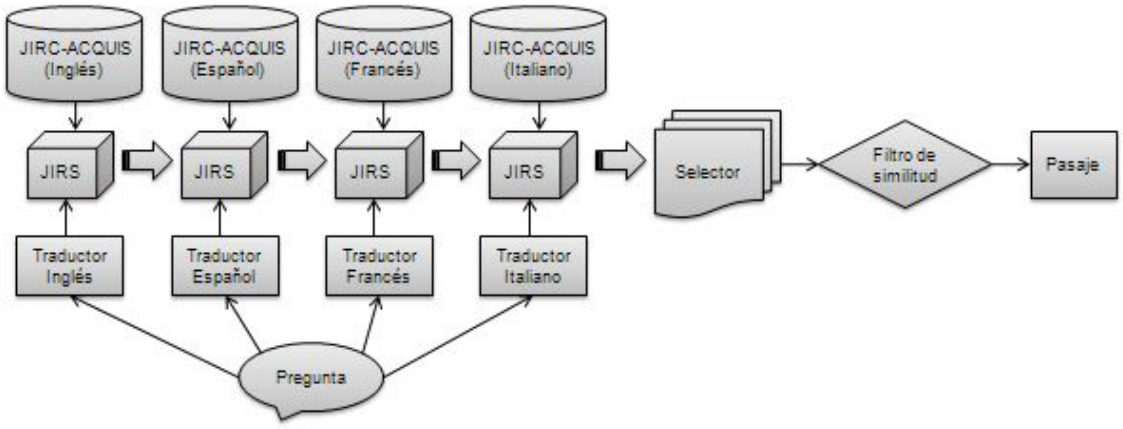

Figura 4.4: Esquema del Sistema de Recuperación de Pasajes para la competencia *ResPubliQA 2009*, enfoque multilingüe

$$p = \max_{x} Sim(P, q|x) \tag{4.3}$$

Donde:

- P: Conjunto de pasajes,
- $p \in P$,
- $x \in$ *Inglés, Español, Francés, Italiano.*

Por lo tanto se propone el esquema de la Figura 4.4. Cada pregunta formulada es traducida, por medio de un sistema de traducción como *Goolge Translator*[6], a los 4 idiomas seleccionados para la implementación del enfoque: Inglés, Español, Francés e Italiano. Posteriormente la pregunta es analizada para determinar su similitud con cada uno de los pasajes de la colección de documentos en los cuatro idiomas, de los cuales se obtiene un pasaje susceptible de ser la respuesta a la pregunta. El módulo *Selector* se encarga de seleccionar como respuesta el pasaje mejor puntuado. Finalmente, el *Filtro de similitud* se encarga de descartar la repuesta si ésta presenta una puntuación de similitud muy baja, de manera similar al filtro aplicado en el enfoque monolingüe.

4.5. Experimentos

En la Tabla 4.2 se aprecian los experimentos realizados empleando los enfoques explicados en la Sección 4.4. Cuatro son las participaciones empleando el enfoque monolingüe que se

[6]http://translate.google.es/#

Tabla 4.2: Distribución de los enfoques aplicados en los experimentos realizados en *ResPubliQA 2009*. LP: Lenguaje en el que se formula la pregunta, LR: Lenguaje en el cual se facilita la repuesta

LP \ LR	Inglés	Francés	Italiano	Español
Inglés	Monolingüe			
Franés		Monolingüe		
Italiano			Monolingüe	
Español				Monolingüe Multilingüe

han presentado en la competencia *ResPubliQA 2009*. En cada una de ellas el lenguaje en el que se formula la pregunta es el mismo en el que se facilita la respuesta (pasaje). Los idiomas seleccionados para realizar los experimentos son: Inglés, Español, Francés e Italiano.

JIRS usa pasajes como unidad básica de datos sobre los cuales se desea calcular la similitud respecto a una pregunta de entrada. En otras palabras la base de datos de *JIRS* está compuesta por pasajes extraídos de varios documentos. Como se explica en la Sección 4.3.1, los documentos suministrados se encuentran codificados en formato *XML*, por lo cual es necesario extraer los pasajes de los documentos y almacenarlos con el formato adecuado para que *JIRS* pueda trabajar con ellos integrandolos a su base de datos. En la Figura 4.5 se aprecia una ejemplo de un documento de la colección y en la Figura 4.6 se aprecia el resultado de procesar el documento para añadirlo a la base de datos de *JIRS* (para más información sobre el formato de archivos que usa *JIRS* ver [26]). Posteriormente el sistema calcula la similitud entre cada una de las preguntas de entrada y los pasajes almacenados en la base de datos de *JIRS*, encontrando según la explicación del enfoque monolingüe (ver Sección 4.4.1), el pasaje con mayor posibilidad de ser una respuesta satisfactoria. En la Tabla 4.3 se muestran algunos de los resultados obtenidos por el sistema para la competencia *ResPubliQA 2009*.

En la competencia se presentó una participación, empleando el enfoque multilingüe. El lenguaje en el que se formulan las preguntas es el mismo lenguaje en el que se facilitan las respuestas (pasajes). En este caso el lenguaje seleccionado es el Español. De manera similar al trabajo realizado en los experimentos monolingües, cuatro sistemas de búsqueda han sido implementados para configurar el sistema planteado en la Sección 4.4.2. Una vez realizado este procedimiento el sistema puede calcular la similitud entre los pasajes y las preguntas, y de esta manera encontrará el pasaje que con mayor probabilidad satisfaga las necesidades del usuario [15].

```
<TEI.2 id="jrc31958R0001-es" n="31958R0001" lang="es">
     <teiHeader lang="en" date.created="2007-04-24">
          <fileDesc>
               <titleStmt>
                    <title>JRC-ACQUIS 31958R0001 Spanish</title>
                    <title>Reglamento nº 1 por el que se fija el régimen lingüístico de la↩
                         Comunidad Económica Europea</title>
               </titleStmt>
          </fileDesc>
</teiHeader>
<text><body>
<head n="1">Reglamento nº 1 por el que se fija el régimen lingüístico de la Comunidad ↩
    Económica Europea</head>
<div type="body">
<p n="2">++++</p>
<p n="3">REGLAMENTO N * 1</p>
<p n="4">por el que se fija el régimen lingueistico de la Comunidad Economica Europea</p>
<p n="5">EL CONSEJO DE LA COMUNIDAD ECONOMICA EUROPEA ,</p>
<p n="6">Visto el articulo 217 del Tratado , segun el cual el régimen linguistico de las ↩
    instituciones de la Comunidad sera fijado por el Consejo , por unanimidad , sin ↩
    perjuicio de las disposiciones previstas en el reglamento del Tribunal de Justicia ,</↩
    p>
...
```

Figura 4.5: Formato XML de los documentos de la colección de datos *JRC-Acquis*

```
<DOC>
     <DOCNO>jrc31958R0001-es.xml:1</DOCNO>
     <TEXT>Reglamento nº 1 por el que se fija el régimen lingüístico de la Comunidad ↩
         Económica Europea</TEXT>
</DOC>
<DOC>
     <DOCNO>jrc31958R0001-es.xml:2</DOCNO>
     <TEXT>++++</TEXT>
</DOC>
<DOC>
     <DOCNO>jrc31958R0001-es.xml:3</DOCNO>
     <TEXT>REGLAMENTO N * 1</TEXT>
</DOC>
<DOC>
     <DOCNO>jrc31958R0001-es.xml:4</DOCNO>
     <TEXT>por el que se fija el régimen lingueistico de la Comunidad Economica Europea</↩
         TEXT>
</DOC>
<DOC>
     <DOCNO>jrc31958R0001-es.xml:5</DOCNO>
     <TEXT>EL CONSEJO DE LA COMUNIDAD ECONOMICA EUROPEA ,</TEXT>
</DOC>
<DOC>
     <DOCNO>jrc31958R0001-es.xml:6</DOCNO>
     <TEXT>Visto el articulo 217 del Tratado , segun el cual el régimen linguistico de las↩
          instituciones de la Comunidad sera fijado por el Consejo , por unanimidad , sin ↩
         perjuicio de las disposiciones previstas en el reglamento del Tribunal de ↩
         Justicia ,</TEXT>
</DOC>
...
```

Figura 4.6: Formato de documento para su indexación a *JIRS*

Tabla 4.3: Ejemplos de resultados de la competenica *ResPubliQA 2009*

Pregunta	Respuesta (Pasaje)
¿Qué es un disco CD-R?	(15) Un CD-R es un disco de policarbonato, recubierto con una solución coloreada, una capa de material reflectante de oro o plata y una capa de protección. Los discos de este tipo sólo se pueden grabar una vez, por lo que se denominan discos de tipo WORM (Write Once Read Many - grabación única lectura múltiple). El disco es un soporte de almacenamiento óptico de datos digitales o música. La grabación se realiza exponiendo la solución coloreada a un rayo láser infrarrojo en un grabador de CD-R.
¿Qué comprenden los gastos de la policía de trafico?	Los gastos de policía de tráfico comprenden el conjunto de gastos de los servicios de policía imputables a la actividad que estos ejercen en beneficio del control y de la fluidez del tráfico, incluidos los gastos de edificios, vehículos y equipos dedicados especialmente a estos servicios.

4.6. Discusión de los resultados

En total, 28 participaciones fueron presentadas por 11 equipos europeos e indios en la competencia *ResPubliQA 2009*. La Tabla 4.4 muestra la participación de los equipos según el idioma de las preguntas y las respuestas (pasajes) [59]. Con los enfoques propuestos en la Sección 4.4 hemos participado en 4 de los 8 lenguajes habilitados en la competencia. Las participaciones desarrolladas en el marco del presente trabajo, son realizadas tanto en tareas de gran interés para una importante cantidad de equipos (tareas en la cuales preguntas y respuestas (pasajes) son formuladas en Inglés o Español con 10 y 6 participaciones respectivamente) así como en tareas en las que no se advierte gran interés (tareas en las cuales preguntas y respuestas (pasajes) son formuladas en Francés o Italiano con 3 y 1 participaciones respectivamente). La sencillez del enfoque presentado y la independencia del idioma inherente a él, facilita una herramienta flexible que con pocas modificaciones permite su adaptación a prácticamente cualquier idioma.

Los enfoques propuestos en la Sección 4.4 participan en *ResPubliQA 2009* bajo el nombre del equipo *NLEL*[7] (Natural Language Engineering Lab). Dichos enfoques son considerados sencillos al ser basados únicamente en el método de n-gramas y comparados con los enfoques presentados por el resto de equipos participantes en la competición. Una amplia variedad de técnicas han sido implementadas para el desarrollo de los sistemas. Entre los equipos que

[7] http://users.dsic.upv.es/grupos/nle/

Tabla 4.4: Participación por idiomas en *ResPubliQA 2009*. LR: Lenguaje Respuestas, LP: Lenguaje Preguntas, EU:Basco, BG: Búlgaro, EN: Inglés, FR: Francés, DE: Alemán, IT: Italiano, PT: Portugués, RO: Runamo, ES: Español, Tomado de [59]

	LR							
LP	BG	DE	EN	ES	FR	IT	PT	RO
BG								
DE		2						
EN			10					
ES				6				
EU			2					
FR					3			
IT						1		
PT								
RO								4

mejor desempeño tuvo en la competencia, está el equipo *UNED*[8] (*Universidad Nacional de Educación a Distancia*), el cual presentó un sistema complejo [71], que emplea: n-gramas, Reconocimiento de Entidades Nombradas, Expresiones Temporales y Expresiones Numéricas. La Tabla 4.5 enseña algunos de los métodos utilizados por los equipos para la implementación de sus sistemas.

Dos medidas-base por competencia, calculadas con base en técnicas de *Recuperación de Información*, han sido suministradas por los organizadores de la competencia, con la finalidad de establecer una comparación de dichas técnicas con sistemas diseñados específicamente para la solución de la tarea. La técnica que mejores resultados proporciona para este tipo de tarea [59] es la *función de ranking Okapi-BM25*[68]; la diferencia entre las dos medidas-base radica en que para una de ellas se aplica un pre-proceso de *stemming* sobre la colección de documentos, para después ser tratado con la técnica *Okapi-BM25*; mientras que para la otra medida-base no se realiza ningún pre-proceso sobre la colección [60]. Como era de esperar, los mejores resultados de las medidas-base son aportados por el enfoque en el que se hace el pre-proceso de *stemming* sobre la colección de documentos. Adicionalmente la medida de *combinación*, igualmente proporcionada por los organizadores de la competencia, indica la proporción de respuestas correctas indicadas por al menos uno de los sistemas participantes, en el caso de una hipotética y perfecta combinación de los sistemas. En las Tablas 4.6, 4.7, 4.8 y 4.9 se muestran los resultados obtenidos por los distintos participantes en las tareas de los idiomas: Inglés, Francés, Italiano y Español respectivamente.

Según los resultados obtenidos en *ResPubliQA 2009*, al analizar los mejores de ellos, y más específicamente los resultados de los equipos *UNED* y *NLEL*, se obtiene una conclusión importante: la comparación entre preguntas y pasajes, por medio de técnicas de n-gramas para determinar la similitud entre ambas y por ende, la posibilidad de que el pasaje responda

[8] http://www.lsi.uned.es/

Tabla 4.5: Métdos usados para la implementación de los sistemas participantes en *ResPubliQA 2009*, Tomado de [59]

Equipo	Chunking	*n*-gramas	R. Entidades Nombradas	Expresiones Temporales	Expresiones Numéricas	Análisis de Dependencias	Fucniones (sub, obj, etc.)sistema	Transformaciones sintácticas	Parsing semántico	Etiquetado Semántico	Representación Lógica	Teorema Prover
SYNA			x	x	x	x	x		x	x		
ICIA [32]			x		x							
NLEL [15]		x										
UAIC [31]			x	x	x							
MIRA [82]			x	x	x							
ILES [50]	x		x		x	x	x	x				
IIIT [7]	x		x		x							
UNED [71]		x	x	x	x							
LOGA [24]				x	x				x	x	x	x

Tabla 4.6: Resultados de la competencia *ResPubliQA 2009* para la tarea en Inglés. *nlel091enen*: Participación del grupo NLEL con enfoque monolingüe, Tomado de [59]

Sistema	c@1	Exactitud	Contestadas		No contestadas			
			Correctas	Incorrectas	Indicadas	Correctas	Incorrectas	Vacias
combinación	0.9	0.9	451	49	0	0	0	0
uned092enen	0.61	0.61	288	184	28	15	12	1
uned091enen	0.6	0.59	282	190	28	15	13	0
nlel091enen	0.58	0.57	287	211	2	0	0	2
uaic092enen	0.54	0.52	243	204	53	18	35	0
base092enen	0.53	0.53	263	236	1	1	0	0
base091enen	0.51	0.51	256	243	1	0	1	0
elix092enen	0.48	0.48	240	260	0	0	0	0
uaic091enen	0.44	0.42	200	253	47	11	36	0
elix091enen	0.42	0.42	211	289	0	0	0	0
syna091enen	0.28	0.28	141	359	0	0	0	0
isik091enen	0.25	0.25	126	374	0	0	0	0
iiit091enen	0.2	0.11	54	37	409	0	11	398
elix092euen	0.18	0.18	91	409	0	0	0	0
elix091euen	0.16	0.16	78	422	0	0	0	0

Tabla 4.7: Resultados de la competencia *ResPubliQA 2009* para la tarea en Francés. *nlel091frfr*: Participación del grupo NLEL con enfoque monolingüe, Tomado de [59]

Sistema	c@1	Exactitud	Contestadas		No contestadas			
			Correctas	Incorrectas	Indicadas	Correctas	Incorrectas	Vacias
combinación	0.69	0.69	343	157	0	0	0	0
base092frfr	0.45	0.45	223	277	0	0	0	0
base091frfr	0.39	0.39	196	302	2	2	0	0
nlel091frfr	0.35	0.35	173	316	11	0	0	11
iles091frfr	0.28	0.28	138	362	0	0	0	0
syna091frfr	0.23	0.23	114	385	1	0	0	1

Tabla 4.8: Resultados de la competencia *ResPubliQA 2009* para la tarea en Italiano. *nlel091itit*: Participación del grupo NLEL con enfoque monolingüe, Tomado de [59]

Sistema	c@1	Exactitud	Contestadas		No contestadas			
			Correctas	Incorrectas	Indicadas	Correctas	Incorrectas	Vacias
combinación	0.61	0.61	307	193	0	0	0	0
nlel091itit	0.51	0.51	256	237	7	0	5	2
base092itit	0.42	0.42	212	288	0	0	0	0
base091itit	0.39	0.39	195	305	0	0	0	0

adecuadamente a la pregunta, es la técnica que mejores resultados aporta a la competencia.

Los resultados obtenidos por el enfoque multilingüe, planteado en la Sección 4.4.2, permiten apreciar una marcada y positiva diferencia del 9 % de *Exactitud* con respecto a los resultados obtenidos por el enfoque monolingüe, igualmente planteado en la Sección 4.4.1. La hipótesis planteada para explicar dicho funcionamiento, se basa en que los documentos de la colección son formulados originalmente en un lenguaje x y las preguntas de la competencia originalmente son formuladas en un lenguaje y. Al introducir el componente de traducción en ambos "textos" (en el caso de la colección de documentos por su definición como tal de colección paralela y en el caso de las preguntas por medio del enfoque multilingüe planteado)

Tabla 4.9: Resultados de la competencia *ResPubliQA 2009* para la tarea en Español. *nlel091eses*: Participación del grupo NLEL con enfoque monolingüe, *nlel092eses*: Participación del grupo NLEL con enfoque multilingüe, Tomado de [59]

Sistema	c@1	Exactitud	Contestadas		No contestadas			
			Correctas	Incorrectas	Indicadas	Correctas	Incorrectas	Vacias
combinación	0.71	0.71	355	145	0	0	0	0
nlel092eses	0.47	0.44	218	248	34	0	0	34
uned091eses	0.41	0.42	195	275	30	13	17	0
uned092eses	0.41	0.41	195	277	28	12	16	0
base092eses	0.4	0.4	199	301	0	0	0	0
nlel091eses	0.35	0.35	173	322	5	0	0	5
base091eses	0.33	0.33	166	334	0	0	0	0
mira091eses	0.32	0.32	161	339	0	0	0	0
mira092eses	0.29	0.29	147	352	1	0	0	1

se modifica su sintaxis y por tanto se afecta la estructura de n-gramas que será analizada según el enfoque propuesto, dando cabida a la posibilidad de encontrar una mayor similitud entre pregunta y pasaje, y suponiendo la posibilidad de que el pasaje sugiera una respuesta válida a la pregunta, si se evalúa en un ambiente multilingüe. El *NLEL* ha sido el **único equipo** que ha aplicado un enfoque multilingüe, explotando el carácter paralelo del corpus suministrado [59].

Capítulo 5

Recuperación de pasajes en patentes multilingües

La aplicación de *Recuperación de Pasajes* en patentes es un enfoque novedoso, con el que se pretende examinar desde el punto de vista del *Procesamiento del Lenguaje Natural* la tarea de *Propiedad Intelectual*. Tradicionalmente dicha tarea es tratada con herramientas complejas de análisis estadístico [70]. El estudio presentado en este capítulo es una primera aproximación a este tipo de tarea.

5.1. Propiedad Intelectual

El concepto de *Propiedad Intelectual* es usado a partir del siglo XIX con la fundación de la *Oficina Federal Suiza de Propiedad Intelectual*[1], aunque su evolución viene dada a partir del siglo XVI, donde la necesidad de proteger el *plagio de ideas* promueve la generación de leyes que protejan a los individuos y aseguren de esta manera un veloz desarrollo de la sociedad [35]. La *Propiedad Intelectual* concede a un individuo derechos patrimoniales de carácter exclusivo por un tiempo determinado, para explotar en forma industrial y comercial una invención o innovación. Dichos derechos facultan al titular de la propiedad, para evitar que otra persona haga uso de ella sin su consentimiento. Al finalizar el tiempo de vigencia de la patente y al no ser renovada, es permitida la explotación de la patente sin necesidad de pagar regalías al titular de la misma. La *Propiedad Intelectual* se puede clasificar en varios tipos, entre ellos: derechos de autor, marcas, patentes de diseño industrial y secretos comerciales [38, 22].

La necesidad de relacionar las técnicas de *Recuperación de Información* desarrolladas a nivel académico y particular con el ámbito empresarial, fomenta la creación de grupos como el *IRF*[2] (*Information Retrieval Facility*). La meta científica de dicho grupo, es promover la

[1] https://www.ige.ch/index.php
[2] http://www.ir-facility.org/

investigación de sistemas de *Recuperación de Información* de colecciones a gran escala de documentos estructurados, documentos semi-estructurados y documentos multimedia. Para ello IRF promueve:

- Competencias abiertas para la evaluación de sistemas de *Recuperación de Información*;
- Simposios y conferencias relacionadas a la *Recuperación de Información*;
- Creación de colecciones de documentos para su uso en la investigación.

En la Sección 5.2 se trata profundamente el tema de entes que promueven la realización de competencias de *Recuperación de Información*. Posteriormente en la Sección 5.3, se estudia el caso de la competencia de *Propiedad Intelectual* organizada por el *IRF*. Las Secciones 5.4 y 5.5 analizan una novedosa propuesta basada en *Recuperación de Pasajes* para la realización de la tarea de *Propiedad Intelectual*, finalmente en la Sección 5.6 se analizan los resultados obtenidos.

5.2. Competiciones sobre patentes

Promover la investigación en tecnologías de acceso a la información es el objetivo de los tres grandes foros que se realizan en tres regiones a nivel mundial, tal y como se ha explicado en la Sección 1.1. Distintas tareas son propuestas por cada foro en el ámbito de la *Propiedad Intelectual*:

El foro asiático, NTCIR, propone las tareas:

1. *PAT-MN*[3]: Minería de patentes [51].
   El objetivo de la tarea de minería de patentes, es la creación de mapas de evolución técnica de un conjunto de trabajos de investigación y patentes. Por ejemplo, un mapa en el que los trabajos de investigación y patentes de EE.UU. y Japón, se clasifican en términos de tecnologías elementales y sus efectos.

2. *PAT-MT*[4]: Traducción de patentes [20].
   Tres sub-tareas son propuestas:

   - Traducción: tarea de traducción del Inglés al Japonés y viceversa, de oraciones y reclamaciones extraídas de documentos de patentes;
   - Evaluación: El objetivo es explorar los métodos de evaluación automática de traducción;

[3]http://www.ls.info.hiroshima-cu.ac.jp/ nanba/ntcir-8/cfp.html
[4]http://www.cl.cs.titech.ac.jp/ fujii/ntc8patmt/

- Recuperación: tarea de *Búsqueda de Documentos*, en una colección de patentes, que puedan invalidar la demanda en un proceso de patente.

El foro americano, *TREC*, propone una tarea[5]:

1. Chemical IR Track: Tarea de *RI* química, que tiene por objetivo el desarrollo y la evaluación de la tecnología de búsqueda en larga escala en documentos de dominio químico, incluyendo publicaciones y patentes. El objetivo es identificar como adaptar los métodos de *RI* actuales, en textos que contienen nombres químicos y fórmulas.

Por último el foro europeo, *CLEF*, propone la tarea de *Propiedad Intelectual*[6] , la cual será estudiada a profundidad en la Sección 5.3.

5.3. Competición de Propiedad Intelectual del *CLEF*

La competición de *Propiedad Intelectual* del *CLEF 2009*, en adelante *CLEF-IP* (por sus siglas en inglés), tiene como propósito fomentar y facilitar la investigación en el área de la *Recuperación de Información* en patentes. La tarea propuesta tiene como objetivo la búsqueda del estado del arte de una patente (*prior art*). Este proceso permite determinar que la invención que pretende ser patentada es novedosa, ya que en la búsqueda del estado del arte se puede constatar que no existen patentes anteriores o publicaciones que describan la invención. La rapidez en el proceso de búsqueda del estado del arte, es un factor determinante para que un usuario pueda patentar una invención, dado que continuamente se publican documentos a nivel mundial los cuales pueden entrar en conflicto con los intereses de un ususario. Al tratarse de una búsqueda de nivel mundial, la tarea se convierte en una tarea multilingüe, lo cual añade un factor de complejidad elevado a la tarea propuesta por la competición. El sistema de patentes permite a los inventores un monopolio del uso de su invención por un período de tiempo determinado; en contraprestación el autor permite la publicación explicativa de la misma.

CLEF-IP permite a los participantes desarrollar sistemas que resuelven una tarea de la vida real, de importante relevancia para el sector universitario y empresarial. Antes de solicitar una patente, el usuario interesado debe realizar una primera búsqueda del estado del arte para determinar si la invención satisface el requisito de novedad; posteriormente una nueva búsqueda es realizada por el ente evaluador, para corroborar la información suministrada por el usuario y de esta manera poder dar continuidad al proceso.

La tarea consiste en: a partir de una colección de patentes, recuperar aquellas que constituyan el estado del arte de una patente determinada. En la Figura 5.1 se aprecia de manera gráfica la finalidad de la tarea. Las entradas del sistema de *RI* son:

[5]http://www.ir-facility.org/research/evaluation/trec-chem-10
[6]http://www.ir-facility.org/research/evaluation/clef-ip-10

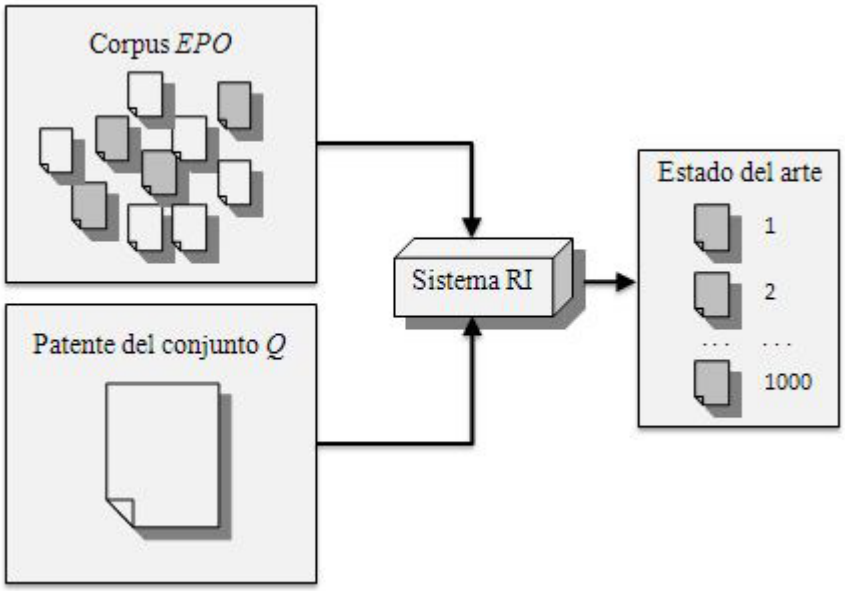

Figura 5.1: Representación gráfica de la tarea de *Propiedad Intelectual* del *CLEF 2009*

Tabla 5.1: Formato de resultados de la competencia *CLEF-IP 2009*. Tomado de [70]

EP1133908	Q0	EP1107664	1	3020
EP1133908	Q0	EP0826302	2	3019
EP1133908	Q0	EP0383071	3	2995
...	...	...	...	...

1. Un conjunto de 500 patentes, que en adelante se denominará conjunto Q, a las cuales se desea determinar el estado del arte individualmente;

2. Una colección de datos, de tamaño considerable que será estudiada en la Sección 5.3.1 y en adelante llamada *corpus EPO*.

La salida del sistema es un conjunto de 1.000 patentes del *corpus EPO*, que conforman el estado del arte de cada una de las patentes del conjunto Q.

Cada una de las 500 respuestas se debe formular según el formato presentado en la Tabla 5.1, en la cual la primera columna corresponde a la patente del conjunto Q analizada; al ser un formato que se rige bajo las normas de la competencia *TREC* descrita en la Sección 5.2, la segunda columna no aporta información relevante para el objetivo de la tarea *CLEF-PI*, pero para no alterar el formato original, se define constante como "*Q0*"; la tercera columna corresponde a la patente extraída del *corpus EPO*, susceptible de hacer parte del estado del arte de la patente analizada del conjunto Q; la cuarta columna es la posición que obtiene la patente extraída entre las 1.000 patentes recuperadas y la quinta columna corresponde a la calificación asignada por el sistema de *RI*.

En la Sección 5.3.1 se describe el *corpus EPO*, posteriormente en la Sección 5.3.2 se describen las medidas utilizadas para calificar las respuestas obtenidas por los equipos participantes.

5.3.1. Corpus de la Oficina de Patentes de la Unión Europea

Una colección de más de 1,6 millones de patentes componen el corpus[7] de la *Oficina Europea de Patentes* (*EPO*[8], por sus siglas en inglés). Dicho ente es el encargado de la aplicación administrativa del *Convenio de Munich*, el cual facilita a una persona de cualquier nacionalidad, solicitar una patente que tenga validez en los países que designe. Entre las funciones que cumple se encuentran: la búsqueda del estado del arte, examen de la novedad, actividad inventiva, aplicabilidad empresarial, suficiencia descriptiva, realización de los examenes de las oposiciones a la concesión de la patente y resolver los recursos interpuestos contra sus actuaciones.

Los registros almacenados en el corpus, datan de documentos publicados entre los años 1978 a 2006, pero sólo se encuentran en formato electrónico, los documentos posteriores al año 1985. Los documentos se proporcionan en formato *XML*; un ejemplo del formato *XML* empleado en los documentos del corpus puede ser apreciado en la Figura 5.2.

En los documentos del corpus *EPO*, se encuentra toda la información referente a una patente y está compuesto por cuatro secciones principales: datos bibliográficos, resumen, descripción y reivindicaciones. Además de estas secciones existen una serie de campos que aportan información adicional tales como: inventor, demandante, citas, entre otras.

Generalmente a una patente están asociados varios documentos publicados, referentes al estado de la patente a lo largo de su proceso de concesión. A cada uno de estos documentos se asigna un código[9] que describe el estado de la patente. Dicho código se designa por una letra seguida generalmente de un dígito que proporciona información adicional sobre la naturaleza del documento. En la Tabla 5.2 se aprecian algunos de los códigos anteriormente descritos. Una patente, por tanto, es considerada un conjunto de documentos; de esta manera puede existir información en el documento *B1* de una patente determinada, que no esté consignada en el documento del tipo *A2* de la misma patente. De igual forma puede existir información del documento *B1* idéntica a la información consiganda en el documento *A2*.

Las patentes de la *EPO* están escritas en uno de sus tres idiomas oficiales: Inglés, Alemán y Francés. El objetivo principal de trabajar con esta colección de datos, es el entrenamiento de sistemas de *RI* capaces de actuar en ambientes multilingües. En la Tabla 5.3 se aprecian algunos datos estadísticos del corpus. Entre ellos se destaca el volumen de datos que deben ser procesados: en total 75 Gb de informació lo cual añade al ejercicio un nivel de complejidad

[7] http://www.epo.org/patents/law/legal-texts.html
[8] http://www.epo.org/
[9] http://www.epo.org/patents/patent-information/raw-data/useful-tables.html

```
   <?xml version="1.0" encoding="UTF-8"?>
<patent-document lang=LANG doc-number=DOCNO date=DATE>
     <bibliographic-data>
          <technical-data status=STATUS>
               <invention-title lang=LANG>INVENTION TITLE</invention-title>
               <citations>
                    <patent-citations>
                         <patcit ucid=UCID source=SOURCE>
                              <document-id format="epo">
                                   <country>COUNTRY</country>
                                   <doc-number>DOCNO</doc-number>
                                   <kind>KIND</kind>
                              </document-id>
                         </patcit>
                    </patent-citations>
               </citations>
          </technical-data>
     </bibliographic-data>
     <abstract lang=LANG>
          <p>ABSTRACT</p>
     </abstract>
     <description lang=LANG>
          <p num="p0001">DESCRIPTION</p>
          ...
          <p num="p000n">DESCRIPTION</p>
     </description>
     <claims lang=LANG>
          <claim id=ID num=NUM>
               <claim-text>
               </claim-text>
          </claim>
     </claims>
</patent-document>
```

Figura 5.2: Formato XML empleado en el corpus *EPO*

Tabla 5.2: Ejemplos de codificación de patentes

A1	Solicitud de Patente (Con reporte de búsqueda)
A2	Solicitud de Patente (Sin reporte de búsqueda)
A3	Patente de Importation
B1	Patente concedida (Con reporte de búsqueda)
B2	Patent concedida (Sin reporte de búsqueda)

Tabla 5.3: Estadísticas de la sub-colección de documentos del corpus *EPO* (Referente a 1.022.388 patentes)

Documentos punlicados entre	1985 y 2000
Documentos publicados	1.958.955
Porcentaje de documentos escritos en Inglés	69 %
Porcentaje de documentos escritos en Alemán	23 %
Porcentaje de documentos escritos en Francés	7 %
Tamaño de la colección	75 Gb

de procesamiento de datos bastante elevado.

5.3.2. Medidas de evaluación

Cuatro medidas de evaluación, fueron utilizadas para determinar el desempeño de los equipos participantes en la competición de *Propiedad Intelectual* del *CLEF* del año 2009 [70]:

1. *Precisión*: definida como el número de documentos relevantes recuperados (n_r), respecto al número de documentos recuperados (N_r).

$$Precision = P = \frac{n_r}{N_r} \tag{5.1}$$

2. *Cobertura*: Se define como, el número de documentos relevantes recuperados (n_r), respecto del número de documentos relevantes en el conjunto de documentos (N_{cd}).

$$Cobertura = C = \frac{n_r}{N_{cd}} \tag{5.2}$$

3. *MAP*: La precisión media *AP* (por sus siglas en inglés), es una estimación global del funcionamiento de un sistema a través de múltiples niveles de cobertura. Para obtener una precisión media de 1.0, el sistema debe recuperar todos los documentos pertinentes (es decir, $Cobertura = 1.0$) y ordenarlos en una lista, de manera tal que documentos irrelevantes no estén mezclados en ella (es decir, $Precision = 1.0$ para todos los documentos listados). El promedio de dichas medidas define el concepto de *MAP* (por sus siglas en inglés).

4. *nDCG*: Medida normalizada de la eficacia de un sistema de *Recuperación de Información*, la cual utiliza una escala de importancia de los documentos clasificados, en un conjunto de resultados devueltos por un motor de búsqueda. nDCG mide la utilidad o ganancia de un documento, basado en su posición en la lista de resultados [33]:

$$nDCG = M \sum_i \frac{(2^{r(i)} - 1)}{log(1 + i)} \tag{5.3}$$

Donde:

- $r(i)$: es la relevancia de los resultados clasificados en la posición i;
- M: es la constante de normalización elegida, para obtener una puntuación entre 0 y 1.

5.4. Sistema de Recuperación de Pasajes para el dominio de patentes

Para solucionar el problema planteado en la Sección 5.3, se parte de una hipótesis simple, en la cual se expone que si se desea encontrar el estado del arte de una patente q en una colección de patentes P, y partiendo del hecho que es posible formular una sentencia (*query*), que resuma adecuadamente el contenido de q, es factible aplicar el *modelo de densidad de distancias de n-gramas* (véase Sección 3.2) en pasajes del conjunto P, para determinar cuáles patentes se encuentran más relacionadas con q, y de esta manera determinar su estado del arte. La baja complejidad del sistema propuesto, dista de propuestas realizadas por otros equipos participantes en *CLEF-IP 2009*, tales como los sistemas implementados por *Lopez y Romary* [42] o el sistema implementado por el grupo *BiTeM* [25] en los cuales se implementan técnicas de ranking tales como *BM25* [68], *Divergencia KL* [79] y *Suavizado de Jelinek-Mercer* [34].

5.4.1. Enfoque multilingue

Como se ha indicado anteriormente en la Sección 5.3.1, la colección de documentos es muy grande y está compuesta por cerca de 75GB de información. Dicha cantidad de información conlleva problemas de memoria al momento de indexar los datos en *JIRS*, por lo cual es necesario realizar una reducción importante para evadir dichos problemas. En la Figura 5.3 se aprecia el esquema del enfoque implementado para la solución de la tarea propuesta por *CLEF-IP 2009*.

La primera reducción que se realiza sobre la colección de patentes, es llevada a cabo por el módulo de eliminación de *Infomación redundante*, así como se explica en la Sección 5.3.1. Básicamente el módulo filtra toda la información pertinente a una misma patente y elimina aquellos documentos que contengan menos información que otros documentos de la misma patente. El criterio de eliminación obedece a la característica de que la mayoría de documentos son actualizaciones de documentos previos; de esta forma se obtiene un único documento por patente. Posteriormene se extraen los campos que consideramos de mayor importancia para la hipótesis postulada por medio del módulo *Extrae campos*. Dichos campos son: título, resumen y descripción. Por último, la colección es separada en tres sub-colecciones, determinadas por cada uno de los idiomas presentes. De esta manera se obtienen tres colecciones de

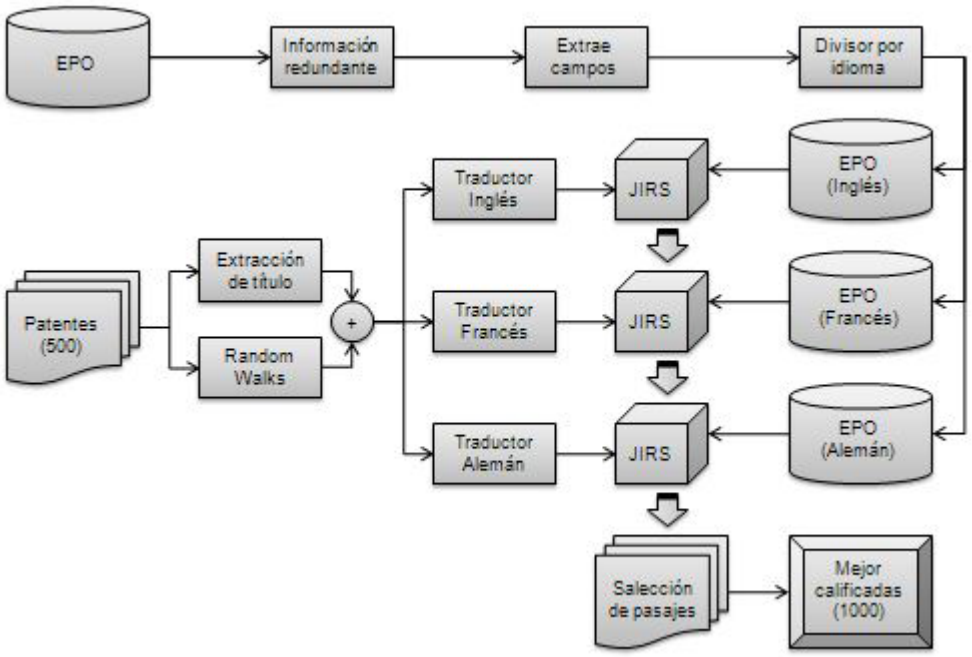

Figura 5.3: Esquema del Sistema de Recuperación del estado del arte de patentes para la competencia de *Propiedad Intelectual del CLEF 2009*, enfoque multilingüe

patentes reducidas que pueden ser indexadas en *JIRS*.

El componente *Patentes (500)* indica un conjunto de 500 patentes que constituyen la entrada al sistema tal y como se indica en la Sección 5.3. Cada una de las patentes de entrada es sometida a un proceso, en el cual se obtiene una sentencia representativa de ella, la cual será usada posteriormente como "query" en el proceso de búsqueda del estado del arte. Dicho proceso está compuesto por los módulos:

- *Extracción de título*: módulo en el cual se extrae el título de la patente en cuestión;
- *Random Walks*: módulo que hace uso de un método de resumen de textos del mismo nombre [30]. Su funcionamiento está inspirado en la teoría de grafos, y usa la co-ocurrencia de términos para asignar una puntuación a una palabra, de acuerdo a su contribución al contexto del documento; de esta manera se extraen las palabras más representativas del campo de *resumen* de la patente.

Las dos cadenas de texto provenientes de cada uno de los dos módulos explicados anteriormente, se encadenan para obtener una única sentencia. Suponga por ejemplo que se desea procesar una patente encontrada en el corpus *EPO* (Véase Sección 5.3.1):

- Número de patente: *EP1127544*

- Título: *Measurement relating to human energy metabolism*
- Descripción: *Aheart rate measurement arrangement chargement comprises a calculating unit comprising a mathematical model arranged to form a person's energy metabolism level as an output parameter of the model unsing as input parameters of the model one or more heart rate parameters and one or more physicological parameters each describing a physiological characteristic of the person...*

Una vez realizado el proceso de *Random Walks* se obtiene la secuencia de palabras: "*the rate heart*"; por lo cual la sentencia que representa a la patente es: "*Measurement relating to human energy metabolism, the rate heart*".

Una vez obtenida la sentencia que representa a la patente en cuestión, tres procesos de traducción[10] operan sobre ella para obtener la sentencia en cada uno de los tres idiomas en los cuales se encuentra la colección de documentos. A continuación se procede a aplicar el *modelo de densidad de distancia de n-gramas* (véase Sección 3.2), con sistemas separados para cada uno de los idiomas. Cada sistema devuelve un listado de patentes con su respectiva calificación, la cual indica la similitud de cada una de ellas con la sentencia de entrada. De estas tres listas se escogen las 1.000 patentes que obtienen la mayor calificación de similitud, para devolverlas como las patentes que conforman el estado del arte (*prior art*) de la patente de entrada.

5.5. Experimentos

Debido al enfoque propuesto y a la cantidad de datos que es necesario procesar para obtener los resultados, sólo un experimento ha sido desarrollado para la competencia *CLEF-IP 2009*. Considerando como entrada al sistema la patente del ejemplo al que se hace referencia en la Sección 5.4.1, la aproximación propuesta devuelve las siguientes patentes:

1.
   - Número de patente: *EP1103216*
   - Título: *Method device measuring blood pressure heart rate environment extreme levels noise vibrations*
   - Descripción: *A method device measuring systolic diastolic blood pressure heart rate environment comprising extreme levels noise vibrations disclosed Blood pressure signals heart beat detected acoustic sensor patient artery...*
2.
   - Número de patente: *EP0785748*
   - Título: *Method and device for determining threshold values for energy metabolism*

[10]La herramienta de traducción utilizada en la implementación es: http://translate.google.com/

Tabla 5.4: Campos usados en la indexación y la formulación de la sentencia (query)

	Indexación				Sentencia (query)				
Gurpo	Título	Recl.	Abs.	Desc.	Título	Recl.	Abs.	Desc.	Otras
clefip-dcu	x	x	x	x	x	x	x	x	-
hcuge	x	-	x	x	x	x	x	x	citas
Hildesheim	x	x	-	-	x	x	-	-	-
humb	x	x	x	x	x	x	x	x	citas, prioridad, aplicaciones, ecla
nlel	x	-	x	x	x	-	x	-	-
clefip-run	-	x	-	-	-	x	-	-	-
Tud	-	x	x	x	x	x	-	-	-
Uaic	x	x	x	x	x	x	x	x	-
clefip-ug	x	x	x	x	x	x	x	x	-
cleipp-unige	x	x	x	-	x	x	x	x	aplicaciones, inventor
UniNE	x	x	x	x	x	x	x	x	-
Utasics	x	x	x	x	x	x	x	x	-

- Descripción: *The invention relates method device determining threshold values energy metabolism method testee subjected gradually increasing stress order obtain threshold values energy metabolism In DE-3439238 disclosed conventional heart rate monitor comprising chest-worn pulse transmitter wrist-worn receiver adapted wirelessly receive pulse signals transmitter...*

Dichos resultados muestarn que existe una relación directa entre las patentes, por lo cual el enfoque presentado relaciona dichas patentes como estado del arte de la patente de entrada.

5.6. Discusión de los resultados

En la Tabla 5.4 se aprecian los campos usados por los equipos participantes en la competencia, para realizar tanto la indexación de las patentes en las bases de datos de sus respectivos sistemas, como los campos empleados para formular las "querys" de las patentes de entrada a los sistemas. En ella se pueden apreciar que equipos que obtienen excelentes resultados como el grupo *hcuge* [25], utilizan casi la totalidad de los campos disponibles. Al contrario, mientras que el equipo con los resultados más bajos no utiliza la totalidad de la información que se suministra en los distintos campos que componen cada patente, lo cual permite concluir que el omitir algunos campos tales como *citas* y *reclamaciones* (véase Sección 5.3.1) conlleva un pobre desempeño de los sistemas ante la tarea propuesta [70].

La omisión de datos tiene una incidencia directa en los resultados. Un ejemplo de ello es que la característica multilingüe de la tarea fue tratada de diversas formas por los distintos equipos: los grupos *Hildesheim* y *Clefip-dcu* [70] usan datos de patentes suministrados en un único lenguaje, por lo que se elimina una gran cantidad de información y por lo tanto se obtien resultados muy bajos. Por otra parte, la mayoría de los equipos, incluido el nuestro [16],

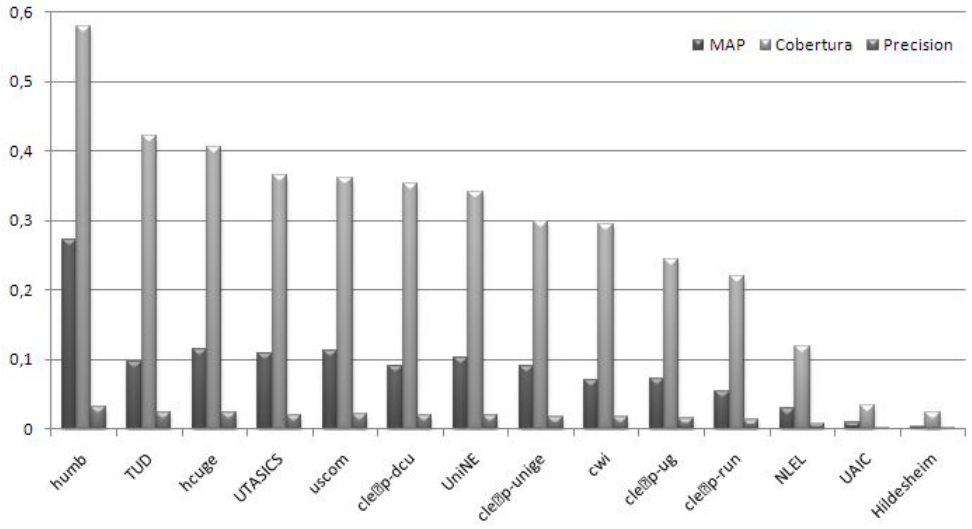

Figura 5.4: Resultados de la competencia de Propiedad Intelectual del *CLEF 2009*: MAP, Cobertura y Precisión

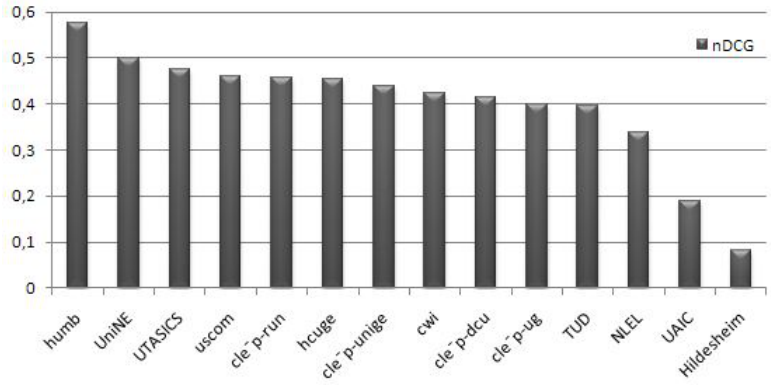

Figura 5.5: Resultados de la competencia de Propiedad Intelectual del *CLEF 2009*: nDCG

usan sistemas de recuperación monolingües para procesar cada uno de los idiomas presentes y posteriormente fusionando los resultados obtenidos.

Sistemas que incorporan conocimientos específicos del procesamiento de patentes tales como los códigos *CIP*[11] (*Clasificación Internacional de Patentes*) ó *ECLA*[12] (*Clasificación Europea de Patentes*), los cuales suministran información de clasificación de las patentes en temas específicos, muestran un mejor desempeño en el desarrollo de la competencia tal y como se aprecia en las Figuras 5.4 y 5.5 [25, 42].

El enfoque propuesto en esta tesis [14], consiste en tratar el problema como un problema de similitud de textos, en el cual la idea base es detectar un posible *plagio de ideas* en

[11]http://www.wipo.int/classifications/ipc/en/
[12]http://www.epo.org/

patentes. La dificultad de la tarea propuesta por *CLEF-IP 2009* generalmente es tratada por medio de sistemas basados en conjuntos de técnicas complejas [70], tal y como se ha visto a lo largo del presente Capítulo. La aproximación implementada por nuestro equipo, plantea la posibilidad de dar solución al problema a través de un sistema simple basado en *Recuperación de Pasajes*.

El pobre desempeño del enfoque propuesto, puede ser debido a la omisión de datos en las distintas fases del procesamiento: en la indexación de documentos se descarta una gran cantidad de información de las patentes compuestas por varios documentos, ya que la eliminación de la información se basó en un criterio de tamaño de los datos en vez de un criterio de comparación de textos, que permitiera extraer la información única contenida en cada documento que compone la patente. La generación de la sentencia (query) de cada una de las patentes de entrada es basada únicamente en los campos *titulo* y *resumen*, lo cual supone la pérdida de una importante cantidad de datos encontrados en los distintos campos de las patentes.

Capítulo 6

Conclusiones y trabajo futuro

La aplicación de enfoques de *Procesamiento de Lenguaje Natural*, en tareas de aplicación potencial en procesos cotidianos que involucran gran cantidad de datos, es el tema central de la presente tesis. Para fundamentar la experimentación realizada, se ofrece una amplia revisión del estado del arte, tanto del procesamiento de textos legales y patentes, como de las técnicas utilizadas para la implementación de los enfoques propuestos a lo largo de la presente investigación. En especial se hace un estudio detallado de la herramienta de *Recuperación de Pasajes JIRS* y su *modelo de densidad de distancias de n-gramas*, en el cual se basan los enfoques planteados.

Se presentan los resultados de los experimentos realizados en el marco de las competencias de *Búsqueda de Respuesta* y *Propiedad Intelectual*, organizadas por el foro *CLEF* en su edición 2009, para evaluar el desempeño de los métodos propuestos para su solución. Dado el carácter multilingüe de las competencias, es deseable y conveniente el uso de sistemas de fácil adaptación para su aplicación en diferentes idiomas, tal es el caso de los enfoques planteados ya que por medio del análisis a nivel de n-gramas, se obtiene una cierta independencia del idioma, la cual permite por medio de pequeños cambios participar en un amplio conjunto de idiomas.

Uno de los aportes realizados en el marco de la presente tesis en competencias de *Búsqueda de Respuesta*, es el uso de un enfoque multilingüe, el cual explota las características de las colecciones de documentos con traducciones paralelas, para la búsqueda de respuestas formuladas en cualquier idioma. Dicho enfoque demostró un mejor desempeño respecto a la aproximación realizada por medio de un enfoque monolingüe, excediendo su calificación de *exactitud* en un 9 %. La hipótesis en la que se basa el buen desempeño del enfoque multilingüe, plantea que existe la posibilidad de encontrar una mayor similitud entre una pregunta redactada en un idioma determinado y un pasaje, si se evalúa en un ambiente multilingüe. Al ser la primera vez que este tipo de aproximación se realiza en competencia, se abre un campo interesante de estudio para futuras herramientas de recuperación de pasajes multilingües.

En cuanto al aporte realizado en el marco de la presente investigación en la competencia de *Propiedad Intelectual*, cabe destacar que se pretende tratar el problema desde un punto de vista de similitud de textos, enfoque que ha sido poco estudiado en este tipo de competencias. La idea base es detectar un posible *plagio de ideas* en patentes, aplicando el *modelo de densidad de distancias de n-gramas*. Los bajos resultados obtenidos en la competencia pueden ser explicados por la omisión de información relevante en el desarrollo de los experimentos.

Actualmente el grupo NLEL participa en la competencia de *Búsqueda de Respuestas* del *CLEF-2010*[1] en la cual además de participar con el enfoque multilingüe propuesto en la edición 2009, se pretenden realizar nuevos estudios con los cuales conseguir un mejor desempeño del sistema.

El pre-procesamiento de la colección de documentos por medio de la técnica de *stemming*, es uno de los estudios con los cuales se pretende mejorar el desempeño de nuestro sistema; la realización de este pre-procesamiento se fundamenta en la investigación de las líneas base de la competencia de *Búsqueda de Respuesta* organizada por el *CLEF* en el año 2009, estudio en el cual se comprueba la efectividad de aplicar esta técnica.

Por otra parte con la aplicación de un filtro basado en la técnica de *Okapi BM25*, se pretende estudiar la posibilidad de re-rankear los 20 pasajes mejor calificados por *JIRS* para cada pregunta, y de esta manera determinar cuál de ellos contiene la respuesta correcta. Las observaciones de estudios sobre la herramienta *JIRS* y participaciones en competencias de *Búsqueda de Respuesta* pasadas, muestran que *JIRS* obtiene un 90 % de efectividad en competencias de *BR* si se le permite devolver dicho número de pasajes.

Una vez sea evaluada la participación en la competencia de *Búsqueda de Respuesta* del *CLEF-2010*, se pretende analizar cuales de los estudios efectuados finalizaron con un mejor desempeño del sistema.

[1] http://clef2010.org/

Referencias

[1] *Eleventh International Conference on Artificial Intelligence and Law, Proceedings of the Conference, June 4-8, 2007, Stanford Law School, Stanford, California, USA*. ACM, 2007.

[2] Mirna Adriani and Rinawati. Finding Answers to Indonesian Questions from English Documents. *Accessing Multilingual Information Repositories*, 4022/2006:510–516, October 2006. DOI: 10.1007/11878773_57.

[3] Gianmaria Ajani, Guido Boella, Leonardo Lesmo, Alessandro Mazzei, and Piercarlo Rossi. Multilingual Ontological Analysis of European Directives. In *ACL '07: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 21–24, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

[4] Carlos Amaral, Adán Cassan, Helena Figueira, André F. T. Martins, Afonso Mendes, Pedro Mendes, José Pina, and Cláudia Pinto. Priberam's Question Answering System in QA@CLEF 2008. In Peters et al. [64], pages 337–344.

[5] Lili Aunimo and Reeta Kuuskoski. Question Answering Experiments for Finnish and French. *Accessing Multilingual Information Repositories*, 4022/2006:477–487, October 2006. DOI: 10.1007/11878773_53.

[6] Sarra El Ayari and Brigitte Grau. A Framework of Evaluation for Question-Answering Systems. In Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soulé-Dupuy, editors, *ECIR*, volume 5478 of *Lecture Notes in Computer Science*, pages 744–748. Springer, 2009.

[7] Rohit Bharadwaj, Surya Ganesh, and Vasudeva Varma. A Naïve Approach for Monolingual Question Answering. In *Working Notes for the CLEF 2009 Workshop*, 30 September - 2 October 2009. Corfu, Greece.

[8] Jon Bing. Performance of Legal Text Retrieval Systems: The Curse of Boole. *Law Library*, 79(2):187–202, 1987.

[9] Mitchell Bowden, Marian Olteanu, Pasin Suriyentrakorn, Thomas D'Silva, and Dan I. Moldovan. Multilingual Question Answering Through Intermediate Translation: LCC's PowerAnswer at QA@CLEF 2007. In Peters et al. [67], pages 273–283.

[10] Davide Buscaldi, Yassine Benajiba, Paolo Rosso, and Emilio Sanchis. Web-Based Anaphora Resolution for the QUASAR Question Answering System. In Peters et al. [67], pages 324–327.

[11] Davide Buscaldi, Paolo Rosso, José Manuel Gómez, and Emilio Sanchis. Answering Questions with an n-gram based Passage Retrieval Engine. *Journal of Intelligent Information Systems*, 1(82):Online First, 2009. DOI: 10.1007/s10844-009-0082-y.

[12] Davide Buscaldi, José Manuel Gómez Soriano, Paolo Rosso, and Emilio Sanchis. N-Gram vs. Keyword-Based Passage Retrieval for Question Answering. In Peters et al. [62], pages 377–384.

[13] Nuria Casellas, Pompeu Casanovas, Joan-Josep Vallbé, Marta Poblet, Mercedes Blázquez, Jesús Contreras, José Manuel López Cobo, and V. Richard Benjamins. Semantic Enhancement for Legal Information Retrieval: Iuriservice Performance. In *ICAIL* [1], pages 49–57.

[14] Santiago Correa, Davide Buscaldi, and Paolo Rosso. NLEL-MAAT at CLEF-IP. In *Working Notes for the CLEF 2009 Workshop*, September 30 - October 2, 2009. Corfu, Greece.

[15] Santiago Correa, Davide Buscaldi, and Paolo Rosso. NLEL-MAAT at CLEF-ResPubliQA. In *Working Notes for the CLEF 2009 Workshop*, September 30 - October 2, 2009. Corfu, Greece.

[16] Santiago Correa, Davide Buscaldi, Paolo Rosso, and Alfonso Rios. Passage Retrieval and Intellectual Property in Legal Texts. In *FLACOS-2009*, Toledo, Spain, September 24-25, 2009.

[17] Luís Costa. Question Answering Beyond CLEF Document Collections. *Evaluation of Multilingual and Multi-modal Information Retrieval*, 4730-2010:405–414, September 2007. DOI: 10.1007/978-3-540-74999-8_48.

[18] Pamela Forner, Anselmo Peñas, Eneko Agirre, Iñaki Alegria, Corina Forascu, Nicolas Moreau, Petya Osenova, Prokopis Prokopidis, Paulo Rocha, Bogdan Sacaleanu, Richard F. E. Sutcliffe, and Erik Tjong Kim Sang. Overview of the Clef 2008 Multilingual Question Answering Track. In Peters et al. [64], pages 262–295.

[19] E. Francesconi, S. Montemagni, W. Peters, and D Tiscornia. *Semantic Processing of Legal Texts*. Lecture Notes in Artificial Intelligence. Springer, 6036.

[20] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. Overview of the Patent Translation Task at the NTCIR-7 Workshop. *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pages 389–400, 2008.

[21] Robert J. Gaizauskas, Mark A. Greenwood, Henk Harkema, Mark Hepple, Horacio Saggion, and Atheesh Sanka. The University of Sheffield's TREC 2005 Q&A Experiments. In Ellen M. Voorhees and Lori P. Buckland, editors, *TREC*, volume Special Publication 500-266. National Institute of Standards and Technology (NIST), 2005.

[22] Roya Ghafele. Perceptions of Intellectual Property: a Review. *Intellectual Property Institute*, 2008.

[23] Danilo Giampiccolo, Pamela Forner, Jesús Herrera, Anselmo Peñas, Christelle Ayache, Corina Forascu, Valentin Jijkoun, Petya Osenova, Paulo Rocha, Bogdan Sacaleanu, and Richard F. E. Sutcliffe. Overview of the CLEF 2007 Multilingual Question Answering Track. In Peters et al. [67], pages 200–236.

[24] Ingo Gloeckner and Bjoern Pelzer. The LogAnswer Project at CLEF 2009. In *Working Notes for the CLEF 2009 Workshop*, 30 September - 2 October, 2009. Corfu, Greece.

[25] Julien Gobeill, Douglas Theodoro, and Patrick Ruch. Exploring a wide Range of simple Pre and Post Processing Strategies for Patent Searching in CLEF IP 2009. In *CLEF 2009 Workshop*, 30 September - 2 October, 2009. Corfu, Greece.

[26] Jose Manuel Gómez. *Recuperación de Pasajes Multilingüe para la Búsqueda de Respuestas*. PhD thesis, Departamento de Sistemas Informáticos y Computación. Universidad Politécnica de Valencia, 2007.

[27] Juna Manuel Gómez, Paolo Rosso, and Emilio Sanchis. Re-ranking of Yahoo snippets with the JIRS Passage Retrieval system. In *Workshop on Cross Lingual Information Access, CLIA-2007, 20th Int. Joint Conf. on Artificial Intelligence, IJCAI-07, Hyderabad, India, January 6-12*, 2007.

[28] Bert F. Green, Jr., Alice K. Wolf, Carol Chomsky, and Kenneth Laughery. Baseball: an automatic question-answerer. In *IRE-AIEE-ACM '61 (Western): Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*, pages 219–224, New York, NY, USA, 1961. ACM.

[29] Sven Hartrumpf. Semantic Decomposition for Question Answering. In Malik Ghallab, Constantine D. Spyropoulos, Nikos Fakotakis, and Nikolaos M. Avouris, editors, *ECAI*, volume 178 of *Frontiers in Artificial Intelligence and Applications*, pages 313–317. IOS Press, 2008.

[30] Samer Hassan, Rada Mihalcea, and Carmen Banea. Random-Walk Term Weighting for Improved Text Classification. In *ICSC '07: Proceedings of the International Conference on Semantic Computing*, pages 242–249, Washington, DC, USA, 2007. IEEE Computer Society.

[31] Adrian Iftene, Diana Trandabat, Ionut Pisto, Alex-Mihai Moruz, Maria Husarciuc, Mihai Sterpu, and Calin Turliuc. Question Answering on English and Romanian Languages. In

Working Notes for the CLEF 2009 Workshop, 30 September - 2 October, 2009. Corfu, Greece.

[32] Radu Ion, Dan Stefanescu, Alexandru Ceausu, Dan Tufis, Elena Irima, and Verginica Barbu-Mititelu. A Trainable Multi-factored QA System. In *Working Notes for the CLEF 2009 Workshop*, 30 September - 2 October, 2009. Corfu, Greece.

[33] Kalervo Järvelin and Jaana Kekäläinen. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[34] Frederick Jelinek and Robert L. Mercer. Interpolated Estimation of Markov Source Parameters from Sparse Data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands: North-Holland, May 1980.

[35] Roberta Rosenthal Kwall. Book Review: Intellectual Property Law and Jewish Law: A Comparative Perspective on Absolutism. *Yale J. Law & Humanites*, 22, 2010.

[36] Leah S. Larkey. A patent search and classification system. In *DL '99: Proceedings of the Fourth ACM Conference on Digital Libraries*, pages 179–187, New York, NY, USA, 1999. ACM.

[37] Dominique Laurent, Patrick Séguéla, and Sophie Nègre. Cross Lingual Question Answering Using QRISTAL for CLEF 2006. In Peters et al. [62], pages 339–350.

[38] Mark A. Lemley. Property, Intellectual Property, and Free Riding. *Texas Law Review*, 83:1031, 2005.

[39] Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli, and Giulia Venturi. NLP-based Ontology Learning from Legal Texts. A Case Study. In Casanovas et al. [1], pages 113–129.

[40] Elizabeth DuRoss Liddy. Anaphora in Natural Language Processing and Information Retrieval. *Inf. Process. Manage.*, 26(1):39–52, 1990.

[41] Xiaoyong Liu and W. Bruce Croft. Passage Retrieval Based on Language Models. In *CIKM*, pages 375–382. ACM, 2002.

[42] Patrice Lopez and Laurent Romary. Multiple Retrieval Models and Regression Models for Prior Art Search. In *CLEF 2009 Workshop*, 30 September - 2 October, 2009. Corfu, Greece.

[43] Bernardo Magnini, Danilo Giampiccolo, Pamela Forner, Christelle Ayache, Valentin Jijkoun, Petya Osenova, Anselmo Peñas, Paulo Rocha, Bogdan Sacaleanu, and Richard F. E. Sutcliffe. Overview of the CLEF 2006 Multilingual Question Answering Track. In Peters et al. [62], pages 223–256.

[44] Bernardo Magnini, Simone Romagnoli, Alessandro Vallin, Jesús Herrera, Anselmo Peñas, Víctor Peinado, Felisa Verdejo, and Maarten de Rijke. The Multiple Language Question Answering Track at CLEF 2003. In Peters et al. [66], pages 471–486.

[45] Bernardo Magnini, Alessandro Vallin, Christelle Ayache, Gregor Erbach, Anselmo Peñas, Maarten de Rijke, Paulo Rocha, Kiril Ivanov Simov, and Richard F. E. Sutcliffe. Overview of the CLEF 2004 Multilingual Question Answering Track. In Peters et al. [63], pages 371–391.

[46] Konstantinos Markellos, Katerina Perdikuri, Penelope Markellou, Spiros Sirmakessis, George Mayritsakis, and Athanasios Tsakalidis. Knowledge Discovery in Patent Databases. In *CIKM '02: Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pages 672–674, New York, NY, USA, 2002. ACM.

[47] K. Tamsin Maxwell and Burkhard Schafer. Concept and Context in Legal Information Retrieval. In *Proceeding of the 2008 Conference on Legal Knowledge and Information Systems*, pages 63–72, Amsterdam, The Netherlands, The Netherlands, 2008. IOS Press.

[48] Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. Automatic detection of arguments in legal texts. In *ICAIL '07: Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230, New York, NY, USA, 2007. ACM.

[49] Diego Mollá and José Luis Vicedo. Question Answering in Restricted Domains: An Overview. *Comput. Linguist.*, 33(1):41–61, 2007.

[50] Véronique Moriceau and Xavier Tannier. FIDJI in ResPubliQA 2009. In *Working Notes for the CLEF 2009 Workshop*, 30 September - 2 October, 2009. Corfu, Greece.

[51] Hidetsugu Nanba, Atsushi Fujii, Makoto Iwayama, and Taiichi Hashimoto. Overview of the Patent Mining Task at the NTCIR-7 Workshop. *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pages 325–332, 2008.

[52] Matteo Negri, Milen Kouylekov, Bernardo Magnini, and Bonaventura Coppola. Reconstructing DIOGENE: ITC-irst at TREC 2006. In Voorhees and Buckland [83].

[53] Matteo Negri, Hristo Tanev, and Bernardo Magnini. Bridging Languages for Question Answering: DIOGENE at CLEF 2003. In Peters et al. [66], pages 501–513.

[54] Günter Neumann and Bogdan Sacaleanu. Experiments on Robust NL Question Interpretation and Multi-layered Document Annotation for a Cross-Language Question/Answering System. In Peters et al. [63], pages 411–422.

[55] Günter Neumann and Bogdan Sacaleanu. DFKI's LT-lab at the CLEF 2005 Multiple Language Question Answering Track. In Peters et al. [65].

[56] Mark Osborn, Tomek Strzalkowski, and Mihnea Marinescu. Evaluating document retrieval in patent database: a preliminary eport. In *CIKM '97: Proceedings of the Sixth International Conference on Information and Knowledge Management*, pages 216–221, New York, NY, USA, 1997. ACM.

[57] Raquel Mochales Palau and Marie-Francine Moens. Argumentation Mining: the Detection, Classification and Structure of Arguments in Text. In *ICAIL*, pages 98–107. ACM, 2009.

[58] Andrés Pedreño. Globalización y Sociedad de la Información: Nuevas Vertientes de Análisis Económico. *Revista económica de Castilla - La Mancha*, 10:311–333, 2007.

[59] Anselmo Peñas, Pamela Forner, Richard Sutcliffe, Álvaro Rodrigo, Corina Forascu, Iñaki Alegria, Danilo Giampiccolo, Nicolas Moreau, and Petya Osenova. Overview of ResPubliQA 2009: Question Answering Evaluation over European Legislation. In *Working Notes for the CLEF 2009 Workshop*, 30 September - 2 October, 2009. Corfu, Greece.

[60] Joaquín Pérez, Guillermo Garrido, Álvaro Rodrigo, Lourdes Araujo, and Anselmo Pe nas. Information Retrieval Baselines for the ResPubliQA Task. In *Working Notes for the CLEF 2009 Workshop*, 30 September - 2 October, 2009. Corfu, Greece.

[61] Laura Perret. A Question Answering System for French. In Peters et al. [63], pages 392–403.

[62] Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber, editors. *Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers*, volume 4730 of *Lecture Notes in Computer Science*. Springer, 2007.

[63] Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, and Bernardo Magnini, editors. *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers*, volume 3491 of *Lecture Notes in Computer Science*. Springer, 2005.

[64] Carol Peters, Thomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J. F. Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Vivien Petras, editors. *Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, volume 5706 of *Lecture Notes in Computer Science*. Springer, 2009.

[65] Carol Peters, Fredric C. Gey, Julio Gonzalo, Henning Müller, Gareth J. F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors. *Accessing Multilingual*

Information Repositories, 6th Workshop of the Cross-Language Evalution Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers, volume 4022 of *Lecture Notes in Computer Science*. Springer, 2006.

[66] Carol Peters, Julio Gonzalo, Martin Braschler, and Michael Kluck, editors. *Comparative Evaluation of Multilingual Information Access Systems, 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers*, volume 3237 of *Lecture Notes in Computer Science*. Springer, 2004.

[67] Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivien Petras, and Diana Santos, editors. *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, volume 5152 of *Lecture Notes in Computer Science*. Springer, 2008.

[68] Stephen E. Robertson and Steve Walker. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In W. Bruce Croft and C. J. van Rijsbergen, editors, *SIGIR*, pages 232–241. ACM/Springer, 1994.

[69] Stephen E. Robertson, Steven Walker, and Micheline Beaulieu. Experimentation as a Way of Life: Okapi at TREC. *Inf. Process. Manage.*, 36(1):95–108, 2000.

[70] Giovanna Roda, John Tait, Florina Piroi, and Veronika Zenz. CLEF-IP 2009: retrieval experiments in the Intellectual Property domain. In *Working Notes for the CLEF 2009 Workshop*, 30 September - 2 October, 2009. Corfu, Greece.

[71] Álvaro Rodrigo, Joaquín Pérez, Anselmo Pe nas, Guillermo Garrido, and Lourdes Araujo. Approaching Question Answering by means of Paragraph Validation. In *Working Notes for the CLEF 2009 Workshop*, October 2009. Corfu, Greece.

[72] Bogdan Sacaleanu, Günter Neumann, and Christian Spurk. DFKI-LT at QA@CLEF 2008. *Evaluating Systems for Multilingual and Multimodal Information Access*, 5706/2009:429–437, September 2009.

[73] José Saias and Paulo Quaresma. A Methodology to Create Legal Ontologies in a Logic Programming Information Retrieval System. In *In Law and the Semantic Web*, pages 185–200, 2003.

[74] G. Salton and C. Buckley. Term-weighting Approaches in Automatic Text Retrieval. *Readings in Information Retrieval*, pages 323–328, 1997.

[75] Amit Singhal. Modern Information Retrieval: A Brief Overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[76] Richard F. E. Sutcliffe, Kieran White, Igal Gabbay, and Michael Mulcahy. Question Answering Using the DLT System at TREC 2006. In Voorhees and Buckland [83].

[77] Hristo Tanev, Milen Kouylekov, Bernardo Magnini, Matteo Negri, and Kiril Ivanov Simov. Exploiting Linguistic Indices and Syntactic Structures for Multilingual Question Answering: ITC-irst at CLEF 2005. *Accessing Multilingual Information Repositories*, 4022/2006:390–399, October 2006. DOI: 10.1007/11878773_44.

[78] Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton. Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 41–47, New York, NY, USA, 2003. ACM.

[79] Takashi Tomokiyo and Matthew Hurst. A Language Model Approach to Keyphrase Extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions*, pages 33–40, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[80] Alessandro Vallin, Bernardo Magnini, Danilo Giampiccolo, Lili Aunimo, Christelle Ayache, Petya Osenova, Anselmo Peñas, Maarten de Rijke, Bogdan Sacaleanu, Diana Santos, and Richard F. E. Sutcliffe. Overview of the CLEF 2005 Multilingual Question Answering Track. In Peters et al. [65], pages 307–331.

[81] Tom M. van Engers, Ron van Gog, and Kamal Sayah. A Case Study on Automated Norm Extraction. In T. Gordon, editor, *Legal Knowledge and Information Systems. Jurix 2004: The Seventeenth Annual Conference*, Frontiers in Artificial Intelligence and Applications, pages 49–58. IOS Press, 2004.

[82] María Teresa Vicente-Díez, Paloma Martínez César de Pablo-Sánchez and, Julián Moreno Schneider, and Marta Garrote Salazar. Are Passages Enough? The MIRACLE Team Participation at QA@CLEF2009. In *Working Notes for the CLEF 2009 Workshop*, 30 September - 2 October, 2009. Corfu, Greece.

[83] Ellen M. Voorhees and Lori P. Buckland, editors. *Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006, Gaithersburg, Maryland, November 14-17, 2006*, volume Special Publication 500-272. National Institute of Standards and Technology (NIST), 2006.

[84] Robert Wilensky. Talking to UNIX in English: An Overview of UC. In *AAAI*, pages 103–106, 1982.

Apéndice A

Aplicación del sistema de recuperación de pasajes en la empresa

Generalmente las empresas trabajan con una gran cantidad de documentos: contratos, reclamaciones, manuales, incidencias, informes técnicos, facturas, noticias, blogs, etc. Debido a la sobrecarga de información que debe ser procesada diariamente, disponer de herramientas como los sistemas de *Recuperación de Pasajes*, con el fin de filtrar la información que resulte pertinente, es una necesidad real. El sistema de Recuperación de Pasajes *JIRS*, gracias a su enfoque simple de n-gramas, está capacitado para procesar cualquier documento que esté escrito en *Lenguaje Natural* y para devolver sólo la información pertinente a las necesidades de la empresa.

Para evaluar el desempeño de *JIRS* en este tipo de tarea, hemos llevado a cabo algunas pruebas aplicándola sobre un caso real; para ello contamos con una base de datos de incidencias reportadas por los ingenieros de una empresa de software, respecto a las herramientas de desarrollo que utilizan diariamente. En la Tabla A.1 pueden ser apreciadas algunas características de dicha colección de documentos.

El ejercicio planteado suponía que una nueva incidencia se presentaba en una aplicación, y que por lo tanto el usuario (ingeniero) deseaba saber si una incidencia similar se había presentado anteriormente, para determinar la solución del problema aprovechando la infor-

Tabla A.1: Características de la base de datos de incidencias de Maat Gknowledge

Número de incidencias	1,759
Languaje	Español
Promedio de palabras por incidencia	47
Formato de las incidencias	Texto en lenguaje natural

Tabla A.2: Pasajes relevantes retornados por *JIRS*, en la base de datos de incidencias de Maat Gknowledge

Ranking del pasaje	Similitud	Pasaje
1	1.0	**No aparecen los campos** seleccionados cuando se están creando formularios en el aplicativo de banca Toledo, intentando crear un formulario en la última pestaña titulada visualización no se listan los campos de dentro de los select que muestra esta plantilla.
2	1.0	Se nos ha presentado la siguiente dificultad con la serialización de la caché. Ha sido sobre la última versión y lo hemos podido reproducir en una aplicación pequeña. Al consultar la clase personas, y listar sus campos, **no aparecen los campos**...
74	0.52	En las publicaciones se están guardando **campos** de otros tipos de publicaciones que **no** son necesarios.
887	0.16	Se nos está presentando un fallo con el gsqlcommand al momento de realizar inserciones con **campos** cuyo valor incluye paréntesis, más específicamente con los que terminan en paréntesis derecho), debido a que se pierde dicho carácter...

mación y el trabajo ya realizado, por otro miembro de la empresa.

Uno de los casos estudiados exponía que el problema en cuestión, se relacionaba con algunos campos de información que no se mostraban en la aplicación, por lo tanto el problema podría ser planteado como pregunta de entrada al sistema de la siguiente manera:

- Q: *No aparecen los campos*

Los resultados devueltos por el sistema y su valor de similitud pueden ser apreciados en la Tabla A.2, en ella es posible apreciar que los pasajes con un valor de similitud alto, son incidencias relevantes.

El enfoque aplicado demostró su capacidad para recuperar un subconjunto de pasajes

pertinentes, basándose en un enfoque simple de n-gramas. El escaso número de pasajes recuperados hace posible, en una segunda fase, el empleo de métodos formales para analizar la información a un nivel más específico.

Apéndice B

Publicaciones

Gracias a las experimentaciones descritas en esta tesis se han publicado los siguientes artículos:

1. Santiago Correa, Davide Buscaldi, and Paolo Rosso. NLEL-MAAT at CLEF-IP. In *Working Notes for the CLEF 2009 Workshop, Corfu, Greece, September 30 - October 2, 2009*, ISSN: 1818-8044.

2. Santiago Correa, Davide Buscaldi, and Paolo Rosso. NLEL-MAAT at CLEF-ResPubliQA. In *Working Notes for the CLEF 2009 Workshop, Corfu, Greece, September 30 - October 2, 2009*, ISSN: 1818-8044.

3. Santiago Correa, Davide Buscaldi, Paolo Rosso, and Alfonso Rios. Passage Retrieval and Intellectual Property in Legal Texts. In *FLACOS-2009, Toledo, Spain, September 24 - 25, 2009*, ISSN: 0806-3036.

4. Santiago Correa, Davide Buscaldi, and Paolo Rosso. NLEL-MAAT at CLEF-IP. *Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers, Lecture Notes in Computer Science*, Springer, 2010 (en prensa).

5. Santiago Correa, Davide Buscaldi, and Paolo Rosso. NLEL-MAAT at CLEF-ResPubliQA. *Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers, Lecture Notes in Computer Science*, Springer, 2010 (en prensa).